Detection and Analysis of Darwinian Selection in the Mitochondrial Genome of Simian Primates.

Thomas Daniel Andrews

September 1998

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University

Declaration

The original work described in this thesis is the result of my efforts alone. The ideas and interpretation presented are my own, although these may have been shaped by discussions with interested parties. Where information presented has not been obtained by myself this fact has been duly attributed. In the case of Chapter 2, which has already been published, the co-authorship of this paper with my supervisors represents their contribution in terms of professional critique, and not for conducting the analyses described.

T. Daniel Andrews

Abstract

This thesis has investigated a potential episode of adaptive evolution that has occurred among specific genes of the mitochondrial genome of simian primates. Prior studies have shown that the cytochrome c oxidase subunit II gene has a higher evolutionary rate along simian primate lineages and that the cytochrome c protein has a greater amino acid replacement rate on the lineage leading to humans. Here, the nucleotide sequences of mitochondrial genes (cytochrome b, cytochrome c oxidase subunit I, and NADH dehydrogenase subunit 2) from various primate species have been obtained, combined with existing sequences present in the international gene databases, and used to investigate further suspected rate increases in simian primate mitochondrial genes. The cytochrome b gene has a higher non-synonymous substitution rate in simian primates than in other mammals. The cytochrome c oxidase subunit I gene also has an increased non-synonymous rate in simian primates, though of a smaller magnitude. A survey of all mitochondrial genes between apes and other mammals revealed three additional genes that may have a similar rate anomaly (ATP synthase subunit 8, NADH dehydrogenase subunits 1 and 5). In contrast, the NADH dehydrogenase subunit 2 gene and other mitochondrial genes were found to have a lack of rate heterogeneity between apes and other mammals, and these sequences have been used to estimate divergence dates between primates and other mammals. This work presents further evidence for a multi-subunit episode of adaptive evolution to have occurred to the electron transport chain of simian primates.

Preface

One of the original goals of the work described in this thesis has been its publication in peer-reviewed journals. Hence, the four main chapters of this thesis which present original work are written in a style acceptable to the journals of the field of molecular evolution. In the following, each of Chapters 2 to 5 describe discrete bodies of work as individual journal articles would do. Chapter 2 has already been accepted for publication in the *Journal of Molecular Evolution*, and it is planned that other chapters may be submitted for publication in the near future. Writing my thesis in this way has been a valuable education in scientific writing, but it has resulted in a minor amount of repetition in the *Materials and Methods* sections between chapters. I apologise to the reader for this inconvenience.

The part of the work described in the following chapters has been previously presented elsewhere:

Andrews, T. D. and Easteal, S. 'Accelerated evolution of the mitochondrial cytochrome b gene of haplorhine primates' Poster, 40th Meeting of the Australian Society of Biochemistry and Molecular Biology, Canberra, 1996.

Andrews, T. D. and Easteal, S. 'Accelerated evolution of mitochondrial genes from haplorhine primates - evidence for the evolution of a biochemical sophistication?' Seminar, 5th International Meeting of the Society of Molecular Biology and Evolution, Garmisch-Partenkirchen, Germany, 1997.

Andrews, T. D., Jermiin, L. S. and Easteal, S. (1998) 'Accelerated evolution of cytochrome b in simian primates: adaptive evolution in concert with other mitochondrial proteins?' J Mol Evol (*in press*).

Acknowledgements

First, I would like to thank my supervisor Dr. Simon Easteal, and my advisers, Drs. Lars Jermiin and Gareth Chelvanayagam for academic comment and guidance throughout my time as a member of the Human Genetics Group. Thanks also to Dr. Gavin Huttley for help with the data analysis in Chapter 4 and advice on all things Macintosh.

Secondly, thanks go to my fellow students, especially Genevieve Herbert, Cheryl Wise, David Betty, Ingrid Jakobsen, Nerida Harley and Andrea Gullock for providing a core of support throughout my time as a member of the E305 laboratory. Also, a big thank-you is necessary for the technical help and cheery tolerance dispensed by Susan Tan and Lynn Croft while I often ungracefully climbed a steep learning curve, and constantly didn't wash-up my glassware.

I want to thank Glenys Noble for sometimes being the only sane person I knew, Kirsten Balding for much-needed encouragement, and the members of the Molecular Genetics group for technical help, the use of their spectrophotometer and cakes at morning tea. Thanks also to Phil Burg, Michael Smith, Dean Kaufmann and Sue Henderson for their help with computer matters.

Thanks to my Mum, Dad and brother Joe, for not being visibly bored by my rather repetitious topics of conversation over the last three and a half years, and for their constant encouragement. Additional thanks to my Mum for coping with a fledgling that still dropped by for food, and Dad for advice imparted along with a palliative, dry Australian red.

Most importantly, thanks to Allison, the other half of my Gestält, for not only putting-up with a science nerd, but also a newly-evolved computer geek.

Contents

Chapter One

General Introduction

While it was Theodosius Dobzhansky who said that nothing in biology makes sense except in the light of evolution, it could perhaps be said that nothing in molecular evolution makes sense except in the light of the neutral theory. The neutral theory of molecular evolution proposed by Motoo Kimura provided a theoretical and mathematical foundation for the study of evolution at the molecular level, and from his theory most current molecular evolutionary research has grown. The neutral theory in essence states that the great majority of evolution at the molecular level is neutral, meaning that all change introduced into genes by mutation is either detrimental or makes no difference (Kimura 1968; Kimura 1983). Furthermore, species adaptation in the sense that it is studied at the macroscopic level occurs due to "pre-adaptation" of individuals within a population (due to genetic polymorphism), and natural selection chooses individuals with the fittest "pre-adapted" phenotype (Kimura 1989). The neutral theory was highly controversial when it was first proposed, and despite a multitude of attempts to find biological examples where neutral evolution does not apply, the neutral theory fundamentally describes the forces of evolution at the molecular level. However, there are rare and interesting departures from neutral evolution, and this is the subject of this thesis.

The first paper presented in the inaugural issue of the journal *Molecular Biology and Evolution* was written by Max Perutz on the subject of protein evolution, more specifically that of haemoglobin (Perutz 1983). Haemoglobin was one of the first intensively studied proteins, and from the results of his study of species variation between haemoglobins, Perutz proposed a theory of protein evolution that has stood the test of

time. Using haemoglobin as an example, Perutz proposed that the evolution of new function in a protein was most often the result of a very small number of amino acid changes. Indeed, Perutz was able to show that most functional differences in haemoglobins between species could be explained as being the result of one to five amino acid differences. As will be discussed in greater detail in the following section, it is now apparent that this kind of acquisition of new function through a small number of amino acid changes in a protein is not unusual, and is perhaps the norm. What this illustrates is that neutrally evolving proteins can very easily, and stochastically, acquire new or modulated function which can confer substantial selective advantage upon the organism in which it has occurred.

Recently, a great amount of interest has been focussed on the potential role that positive natural selection, or adaptive evolution, may have had in shaping the evolution of the function of proteins. At the commencement of the work presented in this thesis, references to positive selection were all but absent from the literature and only a few innovative, and now heavily cited, case studies of adaptive protein evolution existed. Positive selection is now almost a buzz-word and papers on this subject appear in almost every new issue of molecular evolution journals. The purpose of this introduction is to review this rapidly emerging literature, identify trends in this research and present this as a background to the new work that this thesis describes.

Biochemical Studies of Adaptive Protein Evolution

What we now know of adaptive evolution of protein function has grown from two related areas of evolutionary biology. Early studies of the evolution of protein function were biochemical in nature, while the advent of high capacity nucleotide sequencing has allowed a greater emphasis on genetic and statistical studies. The latter is reviewed in the next section, while the biochemical work is covered here. What we know about the biochemical basis of protein evolution is best understood from case studies, and here some of the most prominent examples of adaptive protein evolution are presented. The reader is also directed to the recent excellent review of this subject by Golding and Dean (1998).

Haemoglobins

Haemoglobin is a well understood protein, and biochemical and structural information has been obtained from a large number of species (Perutz 1983). The biochemical subtleties of a large range of species show how haemoglobin has had its function refined for a large variety of evolutionary roles. It is not possible to give a brief review of this subject without omitting important information and missing interesting facts, therefore the reader is referred to reviews of the subject by Perutz (1983) and Clementi *et al* (1994) and here two haemoglobin case studies are presented instead. These are representative of the haemoglobin literature and illustrate how small changes to a protein, achievable through neutral evolution, can alter or modulate its function.

Crocodile Haemoglobin. Crocodiles have an advantage over most of their prey in that they are capable of staying submerged in water for periods of over an hour (Buffetaut 1979). Crocodiles do this in a different fashion to deep-diving mammals in that they do not have large myoglobin stores in their muscle tissue, but possess a haemoglobin that has a unique allosteric binding site for bicarbonate (Bauer *et al* 1981; Bauer and Jelkmann 1977). When crocodiles dive and hold their breath they can not only shut off blood flow to their muscles, they also have haemoglobin that is sensitive to bicarbonate. The presence of bicarbonate in the bloodstream is an indicator of a lack of oxygen, and when bicarbonate levels rise the affinity haemoglobin has for bound oxygen is lowered, and hence more oxygen is released in the crocodile's blood-stream. The truly interesting feature of the bicarbonate effect of crocodile haemoglobin is that it can be transplanted into human haemoglobin with just twelve amino acid changes (Komiyama *et al* 1995). This is even more remarkable in light of the only 68% and 51% amino acid sequence identity that the human α- and β-subunits of haemoglobin share with their respective crocodilian counterparts (Leciercq *et al* 1981).

High-Altitude Haemoglobins. Many birds fly at extremely high altitudes during their annual migration, and studies have been conducted to look for adaptations apparent in the haemoglobins of these birds that fly in such thin air. Comparative studies have looked at the haemoglobins of goose species that fly at low altitudes and high altitudes (high enough to traverse the Himalayas and Andes) (Heibl *et al* 1987; Jessen *et al* 1991). What has been

found is that the high-flying birds have haemoglobins with a higher affinity for oxygen, and this raised affinity can be attributed to just one amino acid difference between the haemoglobins of the low- and high-flying birds. While there are a few complicating factors, such as the higher affinity only occurs in the presence of specific effectors, it is an unambiguous demonstration that small changes to proteins can have large effects.

Both of these examples of small differences between haemoglobins of different species show that advantageous changes in protein function can be achieved through a limited number of amino acid replacements. Certainly, it is possible to imagine a situation where neutral evolutionary drift could accidentally compile the necessary amino acid changes required for a modulation of enzyme function, the result of which could be to give an organism a competitive edge, and hence could lead to the natural selection of the protein.

Stomach Lysozymes

Unlike many bacteria, mammalian species do not possess cellulase enzymes and therefore cannot directly use plant material as a food source. Ruminants (mammalian species which chew their cud), as well as Colobine monkeys (langurs, colobus and other leaf-eating monkeys) and a species of South American bird called the hoatzin, have surmounted this genetic deficit with the evolution of a second stomach, or foregut, where plant material is fermented along with digestive bacteria which leads to it being broken down into utilisable molecules. As would be expected, the bacterial fauna responsible for breaking down plant material regularly move through from the foregut into the main stomach, and these foregut fermenters have evolved a variant lysozyme enzyme (stomach lysozyme) that is active in the stomach, and can digest the peptidoglycan membrane of these bacteria. In breaking down these bacteria which originally grew from the fermented plant material in the foregut, these species are able to maximise the nutritive benefit obtained from the plants they eat (reviewed in Irwin *et al* 1992; Stewart and Wilson 1987).

Non-stomach lysozyme is part of the humoral immune response, and is found in most bodily secretions such as tears and saliva, where it protects against infection through degradation of the extra-cellular peptidoglycan of invading bacteria (Jollès and Jollès 1984).

Stomach lysozymes are very similar to normal secreted lysozyme, except that they have evolved such that they are stable in the low pH environment of the stomach and are resistant to peptidase-mediated degradation (Dobson *et al* 1984). In ruminants and langurs it has been found that the gene encoding lysozyme has been duplicated at least once, and highly specialised ruminants such as the cow and the sheep can have up to as many as ten functional copies of the gene (reviewed in Irwin *et al* 1992). Stewart *et al* (1987) found that stomach lysozymes of cows (ruminants) and langurs (colobine monkeys) appear to have undergone a period of convergent evolution to confer upon normal lysozyme the biochemical properties required for operating in the stomach. Through analysis of the similarities between the stomach lysozymes of the cow and the langur, Stewart and Wilson (1987) proposed that there had been seven convergent amino acid substitutions in each protein which conferred stability at low pH and resistance to peptidase attack. Following this work, Kornegay *et al* (1994) investigated the lysozyme of the foregut fermenting bird, the hoatzin (pronounced "what-seen", see Sears 1994). They found that the hoatzin had a stomach lysozyme similar to that of the mammalian species, and despite these species being separated by over 300 million years, only one gap needed to be inserted in an alignment of the protein sequences of these three stomach lysozymes. The lysozyme gene of the hoatzin has also shown extensive duplication, and in developing the biochemical characteristics of a stomach lysozyme (Kornegay 1996), the hoatzin enzyme has evolved five of the seven amino acids proposed by Stewart and Wilson (1987) as being responsible for adaptation of lysozyme to the stomach environment.

As was found to be the case with haemoglobin, the multiple instances of convergent evolution of stomach lysozymes illustrate how functional modification of an enzyme can be achieved with just a small amount of evolutionary change. The lysozyme genes of mammals have attracted much attention in more theoretical genetic work, and are commonly used as a model case of adaptive evolution in developing new methods of evolutionary inference. Further details of this are examined in the next section.

Human Red/Green Opsins

The opsins are a family of proteins which mediate vision in the retina of most animals (reviewed in Yokoyama 1997). An opsin combined with a chromophore, 11-*cis*-retinal in

humans, make up a visual pigment which is capable of absorbing light of greatly varying wavelength. The actual wavelength at which a visual pigment absorbs light is entirely dependent on the opsin molecule, which interacts with the chromophore to modulate its absorption characteristics. Humans and other higher primates have trichromatic colour vision, and have three different opsins for seeing bright light. These opsins, when part of a visual pigment, absorb light in the blue (410nm), green (532nm) and red (563nm) regions of the visible spectrum (Dartnall *et al* 1983) and hence facilitate colour vision.

The genes which encode the opsins are a multi-gene family with an ancient evolutionary past (reviewed in Yokoyama 1997). Most mammals have only blue and green opsins, and therefore have dichromatic vision. In higher primates a gene duplication has led to red/green vision, and the evolution of this improved colour discrimination has an interesting functional basis. It is apparent that modifications to certain residues in an opsin can drastically change the wavelength at which a visual pigment absorbs light. Analyses of rhodopsin, a highly divergent member of the opsin family which mediates vision in dim light, have shown that a single amino acid causes a huge shift in the absorption maximum of this molecule from 500nm to 380nm (Nathans 1990; Sakmar *et al* 1989). However, this mutant of rhodopsin has not yet been identified in any living organism. The evolution of the red opsin from the green has occurred in a similar way. Yokoyama and Yokoyama (1990) suggested that just three amino acid changes to the green opsin of humans were necessary to achieve the change in absorption maximum required for the green opsin to become a red opsin. Chan *et al* (1992) and Asenjo *et al* (1994) tested this experimentally and found that in total, seven amino acid changes between red and green opsins were required to inter-convert their functions. However, the three changes suggested by Yokoyama and Yokoyama (1990) caused the vast bulk of the spectral shift seen between the proteins, and what is more, each amino acid change was responsible for a discrete and additive change in absorption wavelength. Yokoyama and Yokoyama (1990) also showed that the Mexican cave fish, *Astyanax fasciatus*, has red/green colour vision similar to humans and has convergently evolved red vision from green vision through a series of similar amino acid substitutions.

The evolution of red/green vision in humans has been suggested as an adaptive change in response to changes in the photic environment inhabited by vertebrates. While the ancient ancestors of vertebrates lived in a photic environment dominated by blues and greens (*i. e.* in the sea), modern vertebrates live in more red-oriented environments (in shallow water or on land). The selective advantage that higher primates derive from being able to differentiate red and green has been suggested as being a greater ease of finding red and yellow fruits in green foliage (Mollon 1991). As was the case for haemoglobins and stomach lysozymes, the opsins are another example of how evolutionary adaptation can arise from just a small amount of molecular change.

Other Examples of Adaptive Protein Evolution

Apart from the examples of adaptive protein evolution described above, there is an ever-growing list of proteins where it has been found that their function can be modulated through just a small number of amino acid changes. This phenomenon has attracted the attention of structural biochemists interested in protein engineering, and it has almost become a test of virtuosity to take homologous proteins and transplant the function of one into the other. A particularly notable example of this is the engineering of isocitrate dehydrogenase (IDH) of *E. coli* to accept $NADP^+$ as a co-enzyme instead of NAD^+. Eubacteria differ from eukaryotes in that their IDH preferentially uses $NADP^+$ instead of NAD^+ as a co-factor, and as a consequence are able to utilise acetate as an energy source. Chen *et al* (1995) used x-ray structures of IDH coordinated with $NADP^+$, and the structure of β-isopropylmalate dehydrogenase coordinated with NAD^+, to determine which residues were responsible for co-enzyme specificity in IDH. They found that by changing six key residues in IDH they could change the preference of the *E. coli* enzyme from $NADP^+$ to NAD^+. Similarly, it has been shown that the catalytic activity of lactate dehydrogenase can be engineered to mirror that of malate dehydrogenase via just one amino acid change (Wilks *et al* 1988). Such studies demonstrate once again the ease with which protein function can be modified through a small number of amino acid changes. This kind of change would most certainly be achievable through neutral drift, and if the modified protein confers a

biochemical advantage upon an organism it is foreseeable, that in some cases this could lead to selection.

Biochemical techniques have shown how evolutionary modification of protein function can be detected, but this is a time consuming approach to determining whether adaptive evolution may have occurred. In each of the case studies outlined above, the particular examples of adaptive evolution have been identified only after decades of research, and adaptive selection was only suspected after huge amounts of other experimental evidence had already accumulated. Given the now "post-genomic" world we live in (a term coined by Steven Benner), there are other ways that potential cases of positive selection can be detected from just nucleotide sequence information. While genetic and statistical sequence studies will never replace biochemical experimentation as hard evidence of adaptive evolution, these provide the tools to conduct relatively time-efficient initial analyses of whether positive selection may have acted among a group of homologous genes. In this way, previously unsuspected instances of adaptive evolution have been identified, and this is the subject of the next section of this review.

Statistical and Genetic Inference of Adaptive Evolution

From the current deluge of cases of positive selection reported in the literature, it would appear that it is no longer unusual to suggest that positive selection has influenced the evolutionary history of a particular group of genes. So far, certain trends have become apparent from the reported cases of positive selection, and selection among bacterial defence proteins (Hughes and Yeager 1997), reproductive proteins (Tsaur and Wu 1997; Vacquier *et al* 1997) and viral proteins (Fitch *et al* 1991; Zhang *et al* 1998) has been most prominent. However, these trends may just reflect the curiosities of the researchers pursuing examples of adaptive selection, and efforts are under way to conduct more comprehensive analyses of the scope that positive selection may have using large-scale genome database search algorithms (Endo *et al* 1996).

The methodological challenge of genetic studies aimed at identifying adaptive evolution is how to find adaptive changes in a sea of neutral changes (Perutz 1983). Presently, current methods do not really address this issue, and focus more on identifying increases

in evolutionary rate that may be the result of the selection of a particular gene (Kreitman and Akashi 1995). The commonest method employed to this end has been the measurement of relative abundances of non-synonymous (non-silent or replacement) substitutions to synonymous (silent) substitutions (K_a/K_s ratios) between pairs of nucleotide sequences. Much effort has gone into the development to methods of accurately estimating these quantities, and now a number of powerful methods exist (Goldman and Yang 1994; Li 1993; Wu and Li 1985; Nei and Gojobori 1986, reviewed in Ina 1996), the choice of method used being dependent on the particular set of genes being analysed and the theoretical leanings of the researcher. Other rationales in the detection of positive selection from sequence information are also employed, but fundamentally they are based on the principle of comparing non-synonymous and synonymous change. Innovative maximum-likelihood methods have recently begun to appear that allow sophisticated assessment of positive selection, either through estimation of non-synonymous/synonymous ratios on individual branches of a phylogeny (Yang 1998; Zhang *et al* 1998) or through detection of specific amino acids that demonstrate a positively selected rate (Nielsen and Yang 1998). As in the preceding section, perhaps the best way to get an idea of how this kind of work is done is to look at a couple of good examples of how sequence-based techniques are employed to detect positive selection.

A really good example of the detection of positive selection from sequence information is that found for the abalone sperm lysins. Two proteins present in the acrosomal vesicle of abalone sperm have evolved to become highly specific for the egg of the same abalone species. Gamete recognition systems in marine species are very important in maintaining species integrity and avoiding hybrids. The proteins, lysin and the 18kDa protein (this is what it is called), act together in the process of a sperm fertilising an egg. Lysin is a highly negatively charged protein which acts to dissolve a hole in the vitelline envelope which surrounds an unfertilised egg, while the 18kDa protein promotes fusion of the two gametes. Swanson and Vacquier (1995) have found that these abalone sperm proteins have undergone extensive divergence even between closely related species of abalone, and that it appears that positive selection has been the force driving this divergence. Using the method of Nei and Gojobori (1986), rates of non-synonymous and syn-

onymous nucleotide substitution were compared for the lysin and 18kDa protein genes between species of Californian abalone. Extraordinarily high rates of non-synonymous substitution were found to exist for both proteins, and in many cases these non-synonymous substitution rates were much higher than corresponding synonymous substitution rates (Swanson and Vaquier 1995; Vacquier *et al* 1997). The K_a/K_s ratios found for these proteins are proposed to be the highest yet discovered for any set of full-length nucleotide sequences (Vacquier *et al* 1997). The findings of this study suggest that gamete-gamete specificity in abalone and other marine organisms is of sufficient importance that natural selection favours gamete recognition proteins being adequately divergent so as to avoid hybrid-fertilisation amongst closely related species. Further support for this has been found in the sea urchin gamete recognition protein, bindin, which has also been shown to have undergone positive selection to create greater sequence divergence between species (Metz and Palumbi 1996).

Another study of positive selection that has been conducted amongst the ribonuclease genes from primates, demonstrates the use of comparisons of non-synonymous and synonymous rates of evolution along the internal branches of a phylogeny (Zhang *et al* 1998). When not all of the species in an analysis have undergone a suspected episode of adaptive evolution, the utility of this method is that it can sometimes determine the actual ancestral lineage along which positive selection has occurred. Eosinophil-derived neurotoxin (EDN) and eosinophil cationic protein (ECP) are homologous ribonuclease genes that arose through a tandem gene duplication following the divergence of apes and Old World monkeys from New World Monkeys. EDN and ECP both have ribonuclease activity, but this is lower for ECP, which also has an independent anti-parasitic activity that acts to form holes in the membranes of invading bacteria. The anti-parasitic activity of ECP is thought to be a derived trait that has evolved after the duplication of the ancestral EDN gene. Both EDN and ECP of apes and Old World monkeys have been shown to have very high rates of evolution, although comparisons of these genes between species do not show the non-synonymous substitution rate as being greater than the synonymous rate. However, Zhang *et al* (1998) have shown that along the lineage directly following the duplication of the ancestral EDN/ECP gene and leading to extant ECP genes, there has been a

brief and extreme elevation of the non-synonymous substitution rate, outstripping the synonymous substitution rate by a factor of greater than four. Zhang *et al* identified this short burst of positive selection through prediction of the sequences of the ancestral ECP and EDN genes from extant genes and a knowledge of primate inter-relationships. Hence, between these ancestral sequences both the non-synonymous to synonymous distance ratios and the actual number of non-synonymous and synonymous substitutions could be measured. From these quantities, any periods of adaptive evolution that may have occurred in these genes could be identified. This work by Zhang *et al* (1998), and similar work by Yang (1998), are first steps towards more specific sequence-based techniques of evolutionary inference that dissect-out and identify actual adaptive substitutions in specific lineages from a background of neutral substitutions.

The above examples are representative of the kind of work that is being done, aimed at detecting episodes of positive selection in genes and proteins. While it is not possible to conduct an exhaustive review of other instances of selection, the following is a listing of good examples that I have found but have not mentioned, and the interested reader may care to look up these references. Positive selection of a *Drosophila* reproduction gene (Tsaur and Wu 1997), coordinated adaptive change in defensin proteins (Hughes and Yeager 1997), instances of selection identified in HIV proteins (Nielsen and Yang 1998; Seibert *et al* 1995; Yamaguchi and Gojobori 1997), adaptive evolution to stabilise a co-opted protein domain (Shirai and Go 1997), selection favoring diversity in the major-histocompatibility complex (Hughes and Nei, 1988), rapid evolution of a homeobox regulatory protein (Sutton and Wilkinson 1997), adaptive evolution of alcohol dehydrogenase in *Drosophila* (McDonald and Kreitman 1991), and positive selection of the *jingwei* gene of *Drosophila* (Long and Langley 1993). The last two references are especially interesting as they illustrate how population based measures of evolutionary rate calculated between and within species can be used to detect positive selection.

Biochemistry and Genetics United

Having addressed issues associated with the investigation of adaptive selection separately as being either biochemical or genetic studies, it is appropriate to also briefly describe research that represents a coming together of biochemical and genetic knowledge.

As previously mentioned, the example of adaptive evolution of stomach lysozymes and the large number of lysozyme nucleotide sequences that have been obtained provides an excellent dataset on which to conduct quantitative analyses of the evolution of these genes. Messier and Stewart (1997) have calculated rates of non-synonymous and synonymous substitution between the stomach lysozymes of colobine monkeys and the conventional lysozymes of other primates. Overall, comparisons between the lysozymes of these species groups displayed non-synonymous to synonymous ratios of 2.7 to 3.0, while comparisons of the same quantities between colobine monkeys and between other non-colobine primates gave ratios of 0.53 to 0.90. Such a result is strong confirmation of the adaptive evolution of stomach lysozymes in colobine monkeys. In addition to this, Messier and Stewart (1997) constructed ancestral lysozyme sequences for the same group of primates and were able to show that the episode of positive selection following the evolution of stomach lysozyme was restricted to the lineage immediately preceding the divergence of the various colobine monkey species. Unexpectedly, an episode of positive selection was also identified to have occurred on the lineage leading to apes, and as yet, a biochemical basis for this finding has not been uncovered.

Reconstruction of ancestral sequences has also been of value in furthering our knowledge of the evolution of the primate opsin genes. As discussed above, three important amino acid positions were found to be influential in shifting the absorption maximum of green-sensitive opsins to that of the red-sensitive opsins. Nei *et al* (1997) have used reconstructed sequences to infer how ancestral primate species saw the world. What could be gleaned from this kind of genetic analysis was that the green-sensitive opsin gene evolved from the red-sensitive gene, not the other way around as previously thought. This work also raised another curious and unexpected issue. A large part of the evolutionary history of mammals has been spent nocturnally, yet red/green colour vision appears to have been maintained throughout this time. Hence it has been inferred that the red/green opsins may

have another accessory function that has saved them from functional degeneration, and it has been suggested they have a role in maintaining circadian rhythms.

Both of these genetic studies complement the extensive biochemical work that has previously been carried out, and in both cases the genetic analysis has revealed evolutionary details that were previously unsuspected. Clearly there is great value to be obtained in combining biochemical and genetic techniques to elucidate the evolutionary pasts of genes and proteins that potentially have been subject to recent adaptive evolution. While a number of biochemical studies have now utilised genetic techniques to gain more information about the problem being researched, there has been little use of genetic evidence to initiate detailed biochemical studies, even though interesting results have previously been achieved with this methodology (Jermann *et al* 1995; Szkudlinski *et al* 1996).

The Potential Case of Adaptive Evolution in the Mitochondria of Simian Primates and the Motivation for this Thesis

The work described in the following chapters is an analysis of a potential case of adaptive evolution among a number of proteins that comprise the electron transport chain of simian primates (see Box 1.2). A combination of preliminary biochemical and genetic evidence gave clues that the recent evolution of a number of mitochondrial-encoded genes (see Box 1.1) may not have been entirely neutral. The literature specific to this thesis has been reviewed in the introductions to each of the following chapters, however a brief overview of the most important experimental information that was present at the beginning of this work is given here. First, the results of a study of the interaction of cytochrome c with the cytochrome c oxidase complex (see Box 1.2) showed that the biochemical nature of this interaction was not the same in both the simian primates and other mammals (Osheroff *et al.* 1983). It appears that amino acid changes in both simian cytochrome c and cytochrome c oxidase have occurred to make the interaction of these proteins somewhat species specific. Whether this has been a co-evolutionary or adaptive phenomena is unknown. Genetic studies have shown that certain genes encoded in the mitochondrial genome (see Box 1.1) of primates display altered, or increased evolutionary rates. Adkins and Honeycutt (1994) have shown that the cytochrome c oxidase subunit II gene has un-

dergone a roughly two-fold increase in evolutionary rate in simian primates. The cytochrome b gene has also been shown to have had a much higher evolutionary rate on the lineage leading to humans than on ~~other mammalian lineages~~ those leading to other mammals (Irwin *et al* 1991). An overall survey of the constancy of evolutionary rates of mitochondrial genes has shown that a number of genes have "non-clocklike" rates amongst mammals, and this has been suggested to be a result of increased evolutionary rates on the human lineage (Janke *et al* 1994). As a whole, this biochemical and genetic information warranted further investigation of the evolution of primate mitochondrial genes. Potentially, the genes of the mitochondrial genome that code for various components of the electron transport chain, perhaps in concert with components encoded by the nuclear genome, could have passed through an episode of adaptive evolution that somehow modified its function.

The following four chapters address a number of details associated with the investigation of this matter. Chapter 2 analyses evolutionary rates of primate cytochrome b genes to ascertain if any increase in these is comparable with that of the cytochrome c oxidase subunit II gene. Chapter 3 applies a similar analysis to the cytochrome c oxidase subunit I gene, and also addresses structural matters potentially responsible for any change in the interaction of cytochrome c oxidase with cytochrome c. Chapter 4 analyses evolutionary rates between primates and other mammals for all mitochondrial genes, in an attempt to try to define the scope of genes that may be associated with an episode of adaptive change. Finally, Chapter 5 is somewhat tangential, and looks at how the wide range of evolutionary rates found for primate mitochondrial genes have effected molecular dating of primate divergence times.

Box 1.1 - Mammalian Mitochondria and Their Genomes

In aerobic eukaryotic cells, mitochondria are the intra-cellular organelles that are responsible for the generation of the bulk of adenosine triphosphate (ATP) from adenosine diphosphate (ADP) (Alberts *et al* 1989, chapter 7). Modern mitochondria are thought to be descendants of respiring bacteria that were endocytosed by anaerobic eukaryotes approximately 1.5 billion years ago, when oxygen levels in the Earth's atmosphere began to rise (Cavalier-Smith 1987; Schwartz and Dayhoff 1978; Whatley *et al* 1979). This has been called the "capture" or "endosymbiont hypothesis", and implies that bacterial cells were recruited when aerobic respiration became a more energetically feasible pathway for making ATP. Similar "capture" events have been proposed to be the origin of chloroplasts, these being the result of the endocytosis of photosynthetic bacteria. Genes from these endosymbiotic bacteria have progressively been transferred to the nuclear genome of eukaryotes, and this is shown by the fact that a number of mitochondrial proteins encoded in the nuclear genome of modern organisms display greater similarity with bacterial enzymes than they do with paralogous enzymes from their "host" cell (Gellissen and Michaelis 1987). However, mitochondria still have their own genome, and although it is greatly reduced in size, it possesses the genetic machinery required to express the genes present in it (Grivell 1984). It is not entirely known why mitochondria (or chloroplasts) still have their own genome, but it has been proposed that it may be an evolutionary dead-end (the result of the nuclear and mitochondrial genetic codes diverging) or that the proteins that remain encoded in the mitochondrial genome cannot be properly imported into and assembled in the mitochondria after they have been synthesised in the cytosol (von Heijne 1986).

The mitochondrial genomes from a vast array of eukaryotic organisms have been sequenced, and the structure of the human genome is shown in Figure 1.1. The genome forms a circular, double-stranded molecule of ~~16,569 base pairs~~ 16.6 Kb (in humans) which has two independent origins of replication, one for each strand (termed the heavy (H-) and light (L-) strands) (Anderson *et al* 1981). The mammalian mitochondrial genome contains 13 protein-coding genes, 22 transfer RNA genes and two ribosomal RNA genes. All proteins encoded by the mitochondrial genome form subunits of the mitochondrial electron transport chain (see Box 1.2), while the tRNA and rRNA genes comprise the translation apparatus that allows the protein-coding genes to be expressed (Grivell 1984). The mitochondrial genome also has a non-coding region of around 1.1Kb, called the D-loop or control region, which contains the origins of transcription for both the H- and L-strands (Anderson *et al* 1981).

In an evolutionary context, the mitochondrial genome differs from the nuclear genome in a number of important ways, which has caused it to be long-favoured in studies of molecular evolution. see addendum ~~Apart from having its own genetic code which has diverged from the original universal genetic code (Fox 1987), mitochondrial DNA has been shown to have~~

Box 1.1 - continued

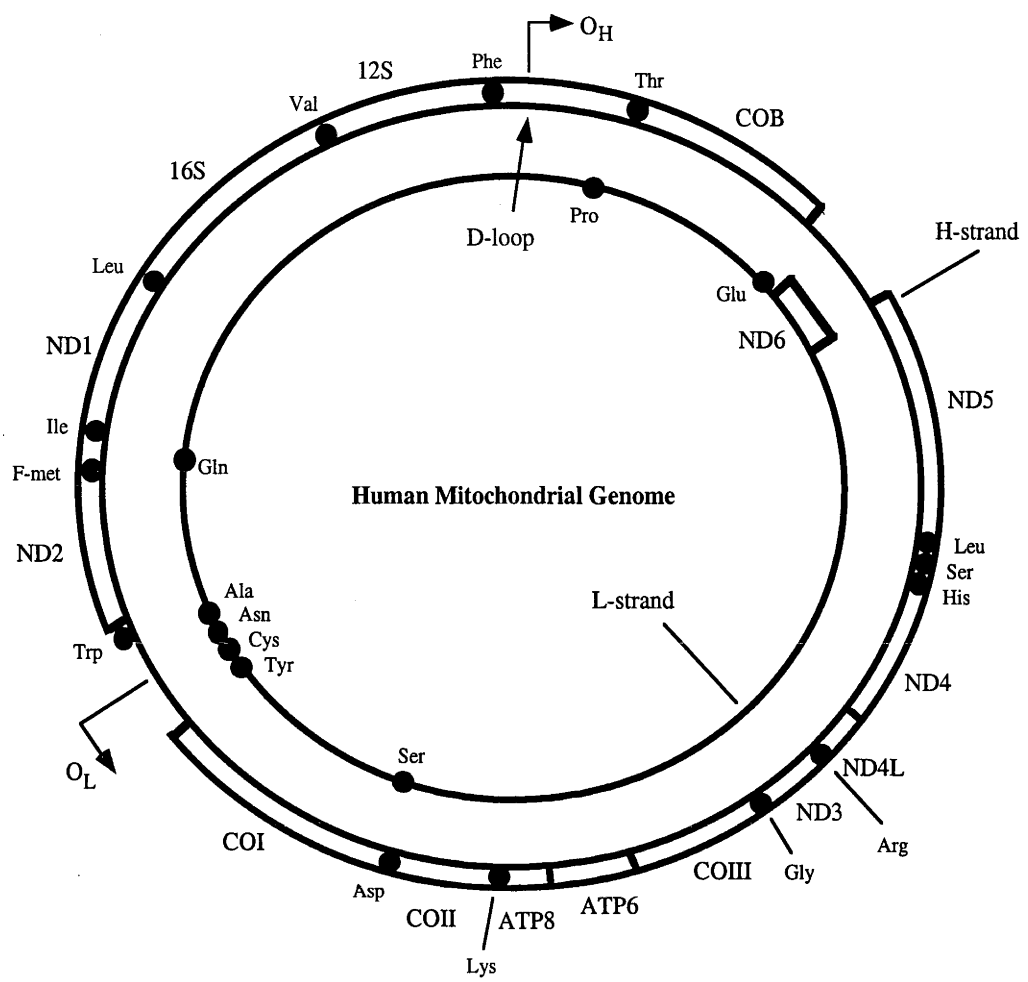


Figure 1.1 - The human mitochondrial genome (Anderson *et al* 1981).

~~an evolutionary rate of up to ten times that of nuclear DNA (Brown 1979).~~ This higher rate of evolution stems from a number of practical matters associated with mitochondrial DNA, such the high error rate of mitochondrial DNA polymerases (Kunkel and Loeb 1981) and the absence of a sophisticated DNA repair mechanism (Clayton 1982; Miquel 1992). Also, mitochondrial DNA is not protected by histones the way nuclear DNA is (Shoffner and Wallace 1992), and the environment in which mitochondrial DNA is stored (the mitochondrial matrix, see Figure 1.2) contains large quantities of free oxidising compounds associated with oxidative phosphorylation (see Box 1.2). Genetic details of mitochondrial DNA are also very different from nuclear DNA. Mitochondrial DNA does not recombine or contain introns and has very few non-coding base pairs, making the mitochondrial genome an example of extreme genetic economy (Attardi 1985), perhaps the result of the large expense that the cell incurs in maintaining the mitochondria's genetic system (von Heijne 1986). Mitochondrial DNA has also been found to be almost entirely maternally inherited (Giles *et al* 1980), and few cases of heteroplasmy in animals have been found (Avise 1986).

Box 1.2 - Mitochondrial Bioenergetics and the Electron Transport Chain

The mammalian electron transport chain (ETC) is comprised of five large protein complexes imbedded in the inner mitochondrial membrane, along with two smaller accessory molecules (Figure 1.2) (Hatefi 1985; Nicholls and Ferguson 1992; Voet and Voet 1990). The ETC catalyses the sequential passage of electrons through a series of flavin, haem, iron-sulfur and copper centres located in the different complexes of the chain. The metabolic benefit of the passage of these electrons is the translocation of protons across the inner mitochondrial membrane to the intermembrane space (Figure 2), which generates a potential gradient which is utilised by the ATP sythase complex to phosphorylate adenosine diphosphate (ADP) to adenosine triphosphate (ATP). The process catalysed by the ETC is the final step in the metabolism of glucose to generate ATP. Glycolysis and the citric acid cycle precede oxidative phosphorylation and metabolise glucose to produce reduced nicotinamide dinucleotide (NADH), and reduced flavin dinucleotide ($FADH_2$) which donate electrons to complex I and ubiquinone of the ETC, respectively. The process of oxidative phosphorylation generates the greatest bulk of ATP production in the cell, producing 32 of the 36 ATP molecules metabolised from each glucose molecule (Voet and Voet 1990).

When electrons from NADH enter the ETC at complex I (also called NADH dehydrogenase or NADH-ubiquinone reductase) they are passed first to a flavin mononucleotide centre within the enzyme complex and then to subsequent iron-sulfur centres, each step representing a drop in the reduction potential of the transported electrons (Weiss *et al* 1991). The passage of two electrons through complex I facilitates translocation of two protons from the matrix to the intermembrane space. Two pairs of electrons are then passed from complex I to ubiquinone (also called coenzyme Q, and ubiquinol when reduced), which is a small, lipid-soluble molecule derived from quinone that can diffuse within the inner mitochondrial membrane and carry electrons between complexes I and III (Vinogradov 1993). Ubiquinone also accepts electrons from the $FADH_2$ centre located within complex II (also called succinate dehydrogenase or succinate-ubiquinone reductase). Complex II is an integral enzyme of the citric acid cycle, and in performing its catalytic function through dehydrogenating succinate to fumarate, the FAD centre of this complex receives two electrons which are passed into the ETC (Singer *et al* 1973; Voet and Voet 1990). Ubiquinol (electron bearing ubiquinone) has a complicated interaction with the cytochrome b subunit of complex III (also known as ubiquinol-cytochrome c reductase or the cytochrome bc_1 complex) which is called the Q-cycle. During the Q-cycle electrons are initially passed to cytochrome b and then back to ubiquinone, and subsequently on to the cytochrome c_1 subunit in complex III (Trumpower 1990; Trumpower and Gennis 1994). This process translocates four protons across the inner mitochondrial membrane, and is an adaptation that allows the four-electron carrying ubiquinone molecule to successfully pass electrons to the two electron carrying cytochrome c_1 (Alberts *et al* 1989). Electrons donated to the ETC by $FADH_2$ in complex II are passed directly to

Box 1.2 -continued

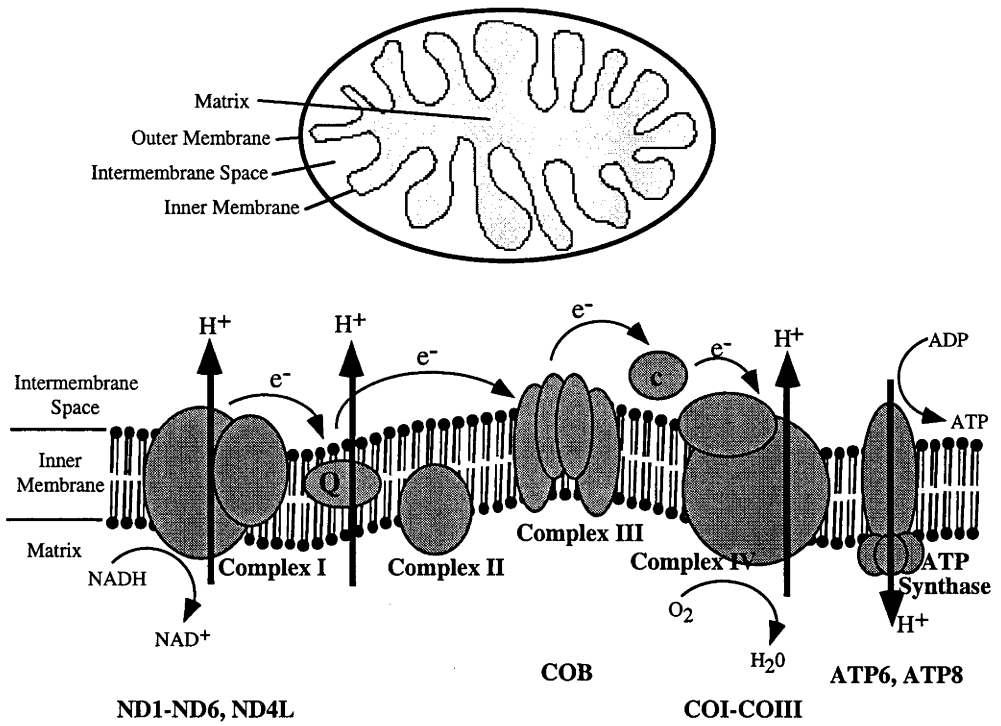


Figure 1.2 - The mitochondria and the mitochondrial electron transport chain.

cytochrome c_1 and do not result in proton translocation. From here, electrons from cytochrome c_1 are passed to the Rieske Iron-Sulphur protein of complex III, which electrostatically interacts with and subsequently passes electrons to membrane-associated cytochrome c (different to cytochrome c_1) which shuttles between complex III and complex IV (most commonly called cytochrome c oxidase) (Capaldi 1990). Electrons from cytochrome c are passed to subunit II of cytochrome c oxidase and then on to subunit I, which together contain haem and copper metal centres and mediate the final step of the ETC in reducing dioxygen to form water (Iwata *et al* 1995; Tsukihara *et al* 1996). This final step in the ETC also results in the translocation of two more protons from the mitochondrial matrix to the intermembrane space.

Electron transport and the generation of a proton gradient is coupled to the process of ADP phosphorylation by the ATP synthase complex (sometimes called complex V) (Penefsky and Cross 1991) and this whole process is termed oxidative phosphorylation. Not all electron transport is coupled to ATP generation, and is used by some animals (particularly those that hibernate during winter) to generate heat (Nicholls and Locke 1984). Generally however, electron transport is tightly coupled to ATP synthesis, and the ETC is inhibited by large proton concentrations in the intermembrane space (Brand and Murphy 1987).

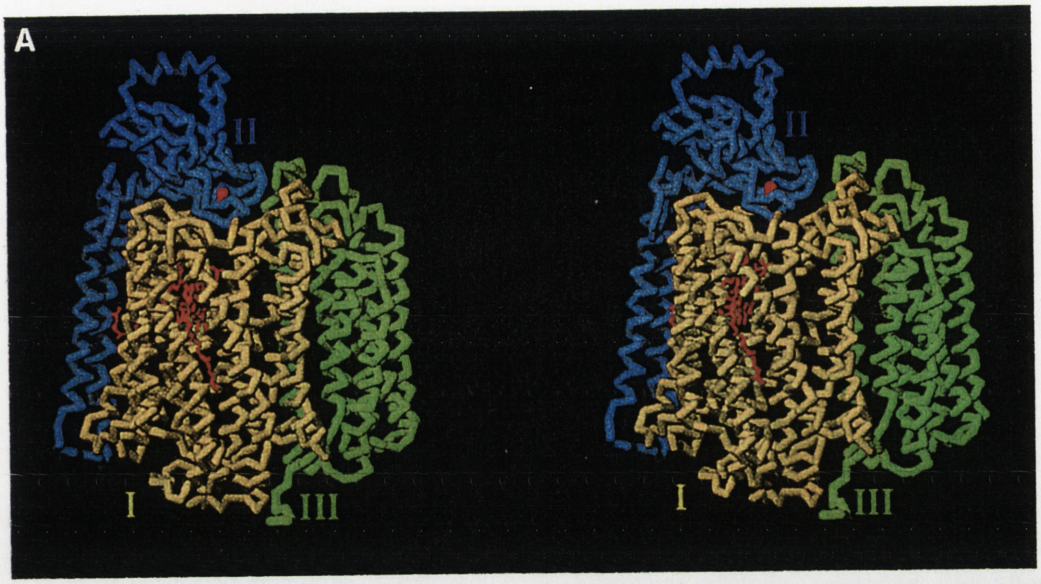
All of the proteins encoded in the mitochondrial genome are components of the ETC, and are present in all complexes with the exception of complex II (see Figure 1.2).

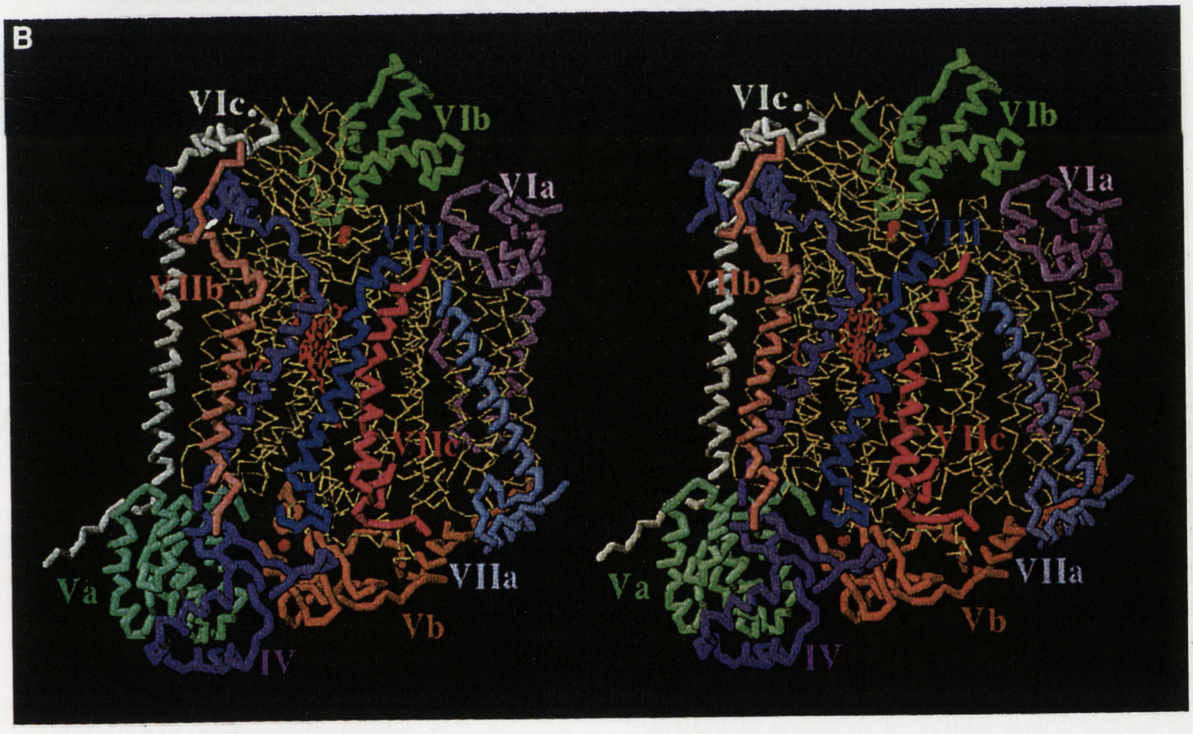
Box 1.2 - continued

Each of the ETC complexes, excluding complex II but including the ATP synthase, are a combination of nuclear- and mitochondrial-encoded subunits, and an issue raised by this thesis is the potential scope for co-evolution that is allowed by the intimate functional contact of these proteins from different genomes. Of greatest interest to this thesis are complexes III and IV, and it is fortunate that during the course of this work crystal structures of these complexes from cow heart were obtained (although the coordinates of complex III were not publicly available at the time when Chapter 2 was written) (Tsukihara *et al* 1996; Xia *et al* 1997). A crystal structure of three subunit cytochrome c oxidase from the bacteria, *Parracoccus denitrificans*, has also been obtained (Iwata *et al* 1995). Representations of the structures of these enormous protein complexes are shown in Figure 1.3. Both complexes from the cow are not only large, but also membrane bound, and their crystallisations were outstanding technical achievements. The structure of complex III shows that the complex is a dimer, consists of two extra-membranous core regions that are structurally similar and confirms the proposed mechanism of the Q-cycle. The structure of cytochrome c oxidase shows the electron transport pathway through the complex, possible channels for proton pumping and a possible site for dioxygen reduction.

Following Pages:

Figure 1.3 - first following page, crystal structure of complex ~~IV~~ III (cytochrome c reductase) showing the position of the cytochrome b subunit relative to the nuclear-encoded subunits (reproduced from Tsukihara *et al* 1996). Second following page, stereo views of complex IV (cytochrome c oxidase, reproduced from Xia *et al* 1997). The top images show the relative positions of the three mitochondrial-encoded subunits (I-III), the bottom images show the full structure, including both mitochondrial- and nuclear-encoded subunits.


A
II
I
III
II
I
III


B
VIc
VIb
VIa
VIII
VIIb
VIIc
VIIa
Va
Vb
IV
VIc
VIb
VIa
VIII
VIIb
VIIc
VIIa
Va
Vb
IV

Chapter Two

Accelerated evolution of cytochrome b in simian primates: adaptive evolution in concert with other mitochondrial proteins?

Abstract

The cytochrome b gene of Horsfield's tarsier, *Tarsius bancanus*, has been sequenced to complete a dataset of sequences for this gene from representatives of each primate infraorder. These primate cytochrome b sequences were combined with those from representatives of three other mammalian orders (cat, whale and rat) in an analysis of relative evolutionary rates. The non-synonymous nucleotide substitution rate of the cytochrome b gene has increased approximately two-fold along lineages leading to simian primates compared to that of the tarsier and other primate and non-primate mammalian species. However, the rate of transversional substitutions at four-fold degenerate sites has remained uniform among all lineages. This increase in the evolutionary rate of cytochrome b is similar in character and magnitude to that described previously for the cytochrome c oxidase subunit II gene. It is proposed that the evolutionary rate increase observed for cytochrome b and cytochrome c oxidase subunit II may underlie an episode of co-adaptive evolution of these two proteins in the mitochondria of simian primates.

Introduction

Analysis of evolutionary rates can be a useful approach for detecting genes and proteins that have undergone an episode of adaptive evolution or have been subjected to positive natural selection. While rates of protein evolution may be increased by elevated mutation rates, this can be distinguished from adaptive evolution through analysis of the relative frequencies of synonymous and non-synonymous nucleotide changes. The hallmark of adaptive evolution is an increased rate of non-synonymous nucleotide change relative to synonymous change, while increased mutation rates result in proportional increases to both synonymous and non-synonymous nucleotide changes (Kreitman and Akashi 1995).

Analysis of the evolutionary rates of seven complete mammalian mitochondrial genomes (rat, mouse, human, seal, cow, whale and opossum as outgroup) has shown significant variation among these lineages for the cytochrome c oxidase subunit I and subunit II genes and the cytochrome b gene (Janke *et al* 1994). More detailed comparative analysis of the cytochrome c oxidase subunit II gene from mammals has confirmed an evolutionary rate acceleration along lineages leading to simian primates (apes including humans, Old and New World monkeys)(Adkins and Honeycutt 1994; Adkins *et al* 1996; Ramharack and Deeley 1987). It appears that cytochrome c oxidase has undergone a nearly two-fold increase in the rate of amino acid substitution relative to other primates (Adkins and Honeycutt 1994), implying an elevated rate of non-synonymous nucleotide substitution. Related to this, the nuclear encoded cytochrome c gene has undergone a period of rapid evolution in the lineage leading to humans (Evans and Scarpulla 1988) and cytochrome c amino acid sequences demonstrate a higher number of replacements on lineages leading to simian primates than they do for other mammalian species (Baba *et al* 1981). The nuclear encoded cytochrome c oxidase subunit IV gene may have also undergone a brief elevation of its non-synonymous substitution rate on ancestral lineages leading to apes and Old World monkeys (Wu *et al* 1997).

Cytochrome c functions by shuttling electrons between the cytochrome c reductase complex (cytochrome b complex) and the cytochrome c oxidase complex and, in doing so, binds directly to subunit II of cytochrome c oxidase (Hatefi 1985). This intimate functional relationship between cytochrome c and cytochrome c oxidase subunit II has led to

the suggestion that their correlated increase in evolutionary rates could be due to co-evolution (Cann *et al* 1984). This suggestion is supported by kinetic studies of *in vitro* reconstituted reactions of cytochrome c with cytochrome c oxidase, which show that the nature of the reaction in simian primates is different from that of other mammals, including strepsirhine primates (Osheroff *et al* 1983).

If adaptive co-evolution has occurred in the mitochondrial electron transport chain of simian primates, the effects of this may not be restricted to cytochrome c and cytochrome c oxidase subunit II. The genes encoding other components of the electron transport chain may also exhibit unusual evolutionary rates as a result of adaptive evolution. While not coming into direct contact with either cytochrome c or cytochrome c oxidase, cytochrome b is a near neighbour, forming an important part of the cytochrome c reductase complex (Hatefi 1985). Cytochrome b has been shown to have an accelerated evolutionary rate on lineages leading to humans (Irwin *et al* 1991; Ma *et al* 1993).

For this study the the cytochrome b gene of *Tarsius bancanus*, has been sequenced, thus completing a dataset representing cytochrome b sequences from each extant primate infra-order. The separation of tarsiers from simian primates is the earliest divergence among extant haplorhine primates (Groves 1989, p105), and hence the tarsier sequence is important in the phylogenetic mapping of any evolutionary rate change that may have occurred among the primates. Here is presented an analysis of the relative rates of evolution of cytochrome b across the primate order in comparison to other mammals.

Materials and Methods

DNA and Data Sources. A lung sample from Horsfield's tarsier (*Tarsius bancanus*) and a liver sample from the Philippine tarsier (*Tarsius syrichta*) were provided by the Duke University Primate Centre. Isolation of pure mitochondrial DNA was precluded by the limited quantity of primary tissue and therefore DNA was phenol/chloroform extracted from homogenates of whole samples (Sambrook *et al* 1989). Other cytochrome b nucleotide sequences were obtained from the following published sources: *Homo sapiens* (human; J01415) (Anderson *et al* 1981), *Colobus guereza* (black-and-white colobus monkey; U38264), *Saimiri sciureus* (squirrel monkey; U38273), *Lemur catta* (ring-tail lemur;

U38271) (Collura and Stewart 1995), *Galago crassicaudatus* (thick-tailed bushbaby; U53579), *Nycticebus coucang* (slow loris; U53580) (Yoder *et al* 1996), *Felis catus* (domestic cat; U20753) (Lopez *et al* 1996), *Balaenoptera physalus* (fin whale; X61145) (Arnason *et al* 1991), *Rattus norvegicus* (rat; X14848) (Gadaleta *et al* 1989), *Giraffa camelopardalis* (giraffe; X56287), *Dama dama* (fallow deer; X56290), *Camelus dromedarius* (dromedary camel; X56281), *Sus scrofa* (pig; X56295), *Equus grevyi* (Grevy's zebra; X56282), *Stenella longirostris* (spinner dolphin; X56292), *Diceros bicornis* (black rhinoceros; X56283) (Irwin *et al* 1991), *Ursus maritimus* (polar bear; X82309), *Phoca groenlandica* (harp seal; X82303) (Arnason *et al* 1995), *Dugong dugong* (dugong; U07564), *Oryctolagus cuniculus* (rabbit; U07566) (Irwin and Arnason 1994), and *Bos taurus* (cattle) (unpublished DDBJ/EMBL/GenBank entry d34635).

PCR and Sequencing. The tarsier cytochrome b gene was isolated by the polymerase chain reaction (PCR) using primers that anneal to conserved regions upstream in the ND6 gene and downstream in the tRNA-Pro gene (Table 2.1 and Figure 2.1). Previously published primers (Irwin *et al* 1991) proved to have a very short lifespan after synthesis, even when stored at -70°C, possibly because they are homologous to tRNA regions and can easily self anneal. PCR products were run on 0.8% low-melting temperature agarose (FMC Bioproducts) gels and the band of the correct size was excised and cleaned using Wizard Preps (Promega). The sequences of both strands of products prepared in this way

Table 2.1 - Oligonucleotide primer sequences.

Primer[a]	Nucleotide Sequence
H14581	5'-CACTAAGGATCCATAAATAGGIGAAGG-3'
H15132	5'-TAGGCTAIGTICTCCCATGAGG-3'
H15359	5'-TCCAITAACCCATCAGGAAT-3'
H15542	5'-CCCCATATTAAICCAGAATGA-3'
L16016	5'-TAGTTTAAAGTAGAAGCTTAGCTTTGGG-3'
L15562	5'-TCATTCTGGITTAATATGGGG-3'
L15379	5'-GATTCCTGATGGGTTAITIGA-3
L15186	5'-TGGCGCCTCAIAATGATATTTG-3'

a - primer numbers refer to the nucleotide position at the 5' end of the oligonucleotide and are numbered after Anderson *et al* (1981).

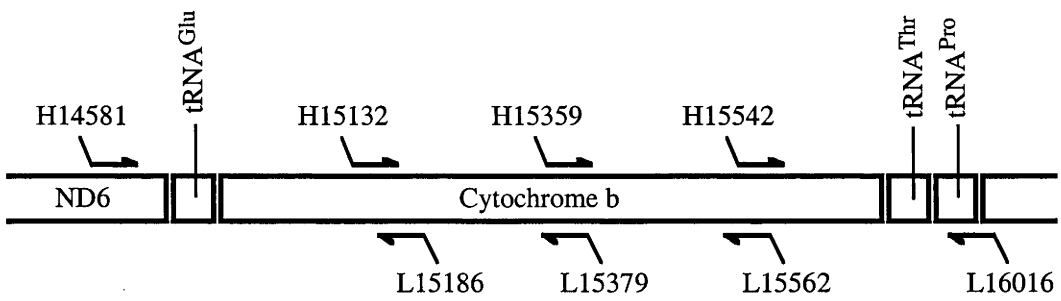


Figure 2.1 - Binding sites of primers (see Table 2.1) used in the amplification and sequencing of the *Tarsius bancanus* cytochrome b gene.

were determined using dye-terminator cycle sequencing (Perkin-Elmer) with the original PCR primers and a series of internal sequencing primers (Table 2.1) on an ABI377 automated sequencer (Applied Biosystems).

Data Analysis. Cytochrome b nucleotide sequences were aligned using CLUSTALW (Thompson *et al* 1994) without the need to insert gaps. Two subsets of the aligned sequences were used. The first ten sequences (*Homo sapiens*, *Colobus guereza, Saimiri scuireus*, *Tarsius bancanus*, *Lemur catta*, *Galago crassicaudatus*, *Nycticebus coucang*, *Felis catus*, *Balaenoptera physalus* and *Rattus norvegicus*) were used to perform an exhaustive search of tree space, to generate maximum likelihood trees and conduct relative rate tests (see below). The second subset of sequences was used to analyse the distribution of variation along the length of an alignment that consisted of nineteen mammalian species, these being representatives of mammalian orders and sub-orders for which complete cytochrome b nucleotide sequences are present in the DDBJ/EMBL/GenBank database. The sequences are those listed in the *DNA and Data Sources* section, including the three strepsirhine primate species and *Tarsius bancanus*, but excluding *Homo sapiens*, *Colobus guereza* and *Saimiri scuireus*.

The number of synonymous (K_s) and non-synonymous (K_a) substitutions per site were calculated using the method of Wu and Li (1985), as modified by Li (1993) and Pamilo and Bianchi (1993) and implemented by the NewDiverge program from the GCG package (Genetics Computer Group, WI, USA).

Potential base compositional heterogeneity among the nucleotide sequences was assessed using the DNA-distance program of Jermiin *et al* (1998), whereby pairwise comparisons of nucleotide sequences were conducted and χ^2 values calculated to test the independence of each pair of sequences with respect to base composition. Such calculations were conducted for datasets consisting of first, second and third codon positions and Z-scores were calculated from the resultant χ^2 values (Smith 1986). The *Z*-scores calculated for each codon position were then graphed as a frequency distribution and used to identify the most compositionally homogenous sites in the data.

The exhaustive search of tree-space for the maximum likelihood tree of the alignment of ten nucleotide sequences was made using the TrExML program (Wolf *et al* 1998) and a standardised, exponentially-weighted majority-rule consensus tree was generated using the TreeCons program (Jermiin *et al* 1997). Relative likelihood support for internal edges in the consensus tree was compared with that for internal edges in the maximum-likelihood tree and the tree representing a conventional primate phylogeny.

Substitution rate estimates between sequences were made with the two-parameter model (Kimura, 1980) and relative rate tests were performedat non-degenerate and four-fold degenerate sites with the method of Wu and Li (1985). *Z*-scores were calculated using the equation given Muse and Weir (1992) and implemented by the K_2WuLi program by L. S. Jermiin. K_{01}:K_{02} values were calculated by with the formulas described by Li (1997, p. 217).

The spatial distribution of variability along the length of the mammalian cytochrome b protein was determined with the Sliding Window program by TDA. A sliding window of twenty residues was moved along the length of the alignment of nineteen mammalian sequences and variability was scored as the number of variable positions of the possible twenty.

Ancestral amino acid sequences were predicted with the maximum-likelihood method of Yang *et al* (1995) and implemented through the PAML package (Yang 1997).

Results and Discussion

Tarsier Cytochrome b

The nucleotide sequence of cytochrome b from Horsfield's tarsier (*Tarsius bancanus*) is 1140 base pairs in length and translates to a protein of 379 residues. In this respect it is like all other mammalian cytochrome b genes sequenced to date, with the exception of the gene from the elephant (Irwin *et al* 1991). The sequence has been submitted to the DDBJ/ GenBank/EMBL databases and has the accession number AB011077.

Portions of a number of independent *Tarsius bancanus* cytochrome b PCR products were sequenced, and while the sequence obtained was identical in each case, varying regions were deleted. To investigate the possibility that a nuclear pseudogene had been preferentially amplified instead of the functional cytochrome b gene additional sequencing of a short 350 base pair region from the product of an identical PCR from *Tarsius syrichta* was performed. No deletions were found in this short region of the *Tarsius syrichta* gene, and a pairwise comparison between the *Tarsius bancanus* and the *Tarsius syrichta* sequences revealed K_a and K_s values of 0.048 and 0.630 substitutions per site, respectively. These values are not characteristic of comparisons between diverged pseudogenes or between a pseudogene and a functional gene. Additionally, calculation of K_s/K_a ratios for comparisons of these sequences with those of other non-simian primates and other mammals resulted in values with an average of 13.19 and a standard deviation of 1.30. Hence, it was concluded that the *Tarsius bancanus* sequence obtained was from a functional copy of cytochrome b. The mitochondrial genome becomes progressively damaged with age, usually due to extensive deletion of semi-random segments (Linnane et al 1992), and hence the deletions found in cytochrome b can perhaps be explained as being the result of damage to the "functional" gene of an aged animal. This is a plausible explanation as our source of DNA was the Duke University Primate Centre, where tissue samples are obtained predominantly from aged animals that have died of natural causes.

Analysis of DNA

Heterogeneity of nucleotide composition has been observed among the cytochrome b genes of mammals (Jermiin *et al* 1994), and this may affect both evolutionary rates and their estimation. Therefore an analysis of evolutionary rates among these genes must take steps to ensure that any results obtained are not compromised by this factor. Table 2.2 shows the base composition at each codon site of each gene analysed in this study. At first and second codon positions the base composition is similar among the sequences, but at third codon positions the variation in base composition is more pronounced. The distribution of base compositional heterogeneity among codon positions can be visualised by plotting *Z*-scores (Smith 1986) for pairwise sequence comparisons of each codon position over the length of the gene (Figure 2.2). The magnitude and range of the *Z*-scores are lower and tighter for first and second codon positions than for third codon positions. This implies a more homogeneous base composition among sequences at the first and second codon positions than among those at the third codon position. Hence, the first and second codon positions provide a better dataset for phylogenetic analysis and estimation of evolutionary rates than does the third codon position.

Table 2.2 - Base compositions of cytochrome b genes calculated for each codon position.

Species	First Codon Position				Second Codon Position				Third Codon Position			
	% A	% T	% G	% C	% A	% T	% G	% C	% A	% T	% G	% C
Homo sapiens	29.6	23.5	19.3	27.7	20.5	40.1	12.9	26.9	36.4	12.1	3.7	47.8
Colobus guereza	31.7	23.2	17.7	27.4	19.8	40.1	12.7	27.4	39.3	18.0	3.7	39.0
Saimiri sciureus	32.7	21.6	18.7	26.9	20.1	39.6	12.7	27.7	38.3	21.6	4.7	35.4
Tarsius bancanus	30.6	23.2	20.8	25.3	20.3	42.0	13.5	24.2	41.2	17.2	1.6	40.1
Lemur catta	30.2	24.9	21.2	23.8	19.8	41.3	14.0	24.7	38.1	23.5	2.9	35.5
Galago crassicaudatus	28.3	23.0	20.9	27.8	20.6	40.2	13.2	26.0	34.9	16.1	4.0	45.0
Nycticebus coucang	28.0	23.2	22.4	26.4	20.1	41.2	13.5	25.3	39.1	19.5	4.5	36.9
Felis catus	27.7	23.7	22.2	26.4	20.1	40.9	14.2	24.8	40.1	15.8	4.5	39.6
Balaenoptera physalus	28.8	21.4	22.7	27.2	20.1	41.7	13.7	24.5	41.4	13.7	2.6	42.2
Rattus norvegicus	28.2	24.0	21.4	26.4	20.1	42.2	14.0	23.7	42.2	15.6	2.6	39.6

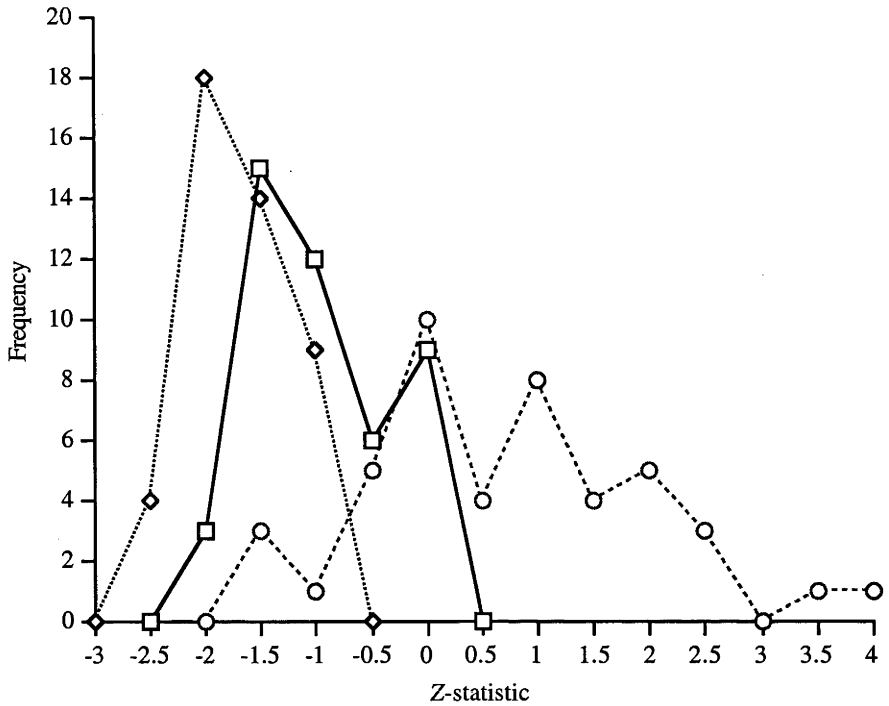
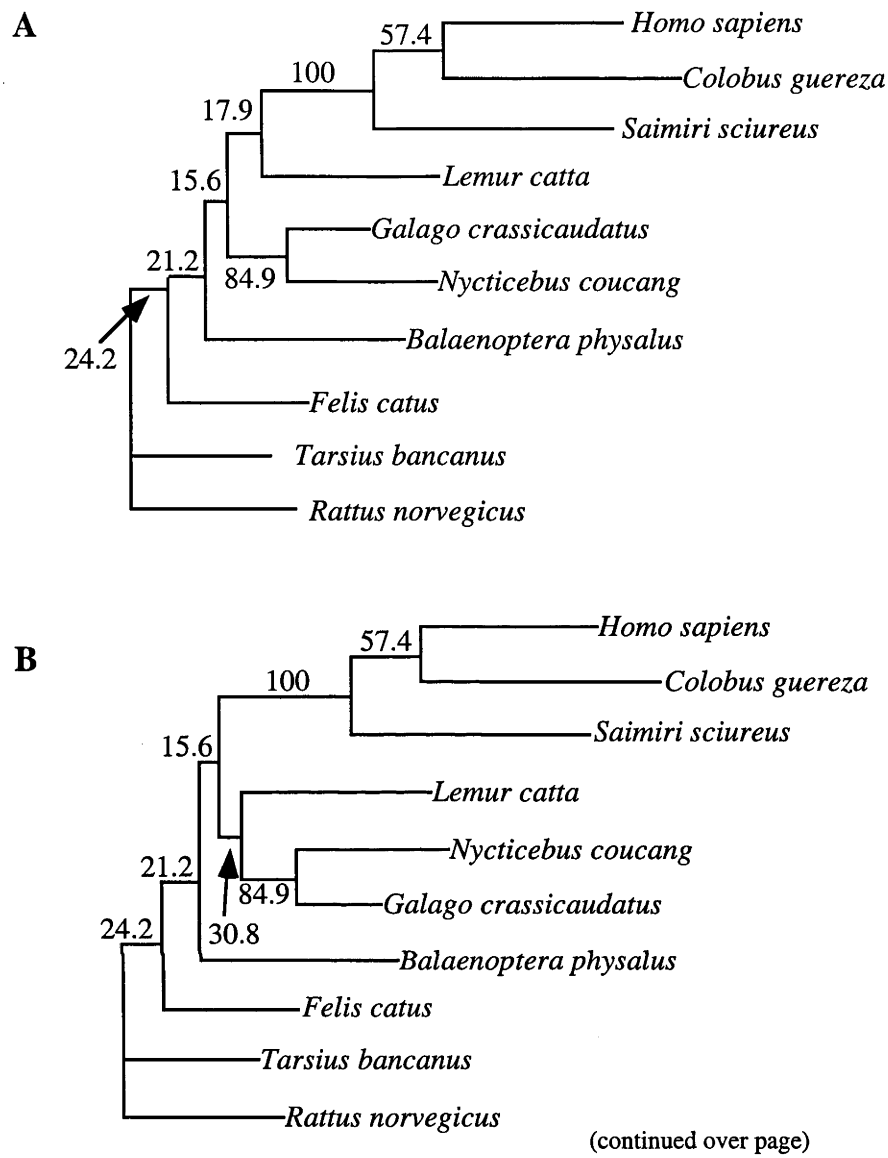


Figure 2.2 - Frequency distribution of Z-stattistics for pairwise comparisons between cytochrome b sequences. Squares show first codon sites; diamonds show second codon sites; and circles, third codon sites.

Tree Generation

An exhaustive search of tree-space (Wolf *et al* 1998) by maximum likelihood analysis was conducted using the alignment of the first and second codon positions of the ten mammalian cytochrome b nucleotide sequences as described under *Materials and Methods*. This analysis produced 1731 trees that were not significantly ($\alpha = 0.05$) different from the maximum likelihood tree shown in Figure 2.3a. A consensus of the 1731 trees and the maximum likelihood tree is shown in Figure 2.3b. The trees shown in Figure 2.3a and 2.3b are congruent neither with each other nor with any generally accepted phylogeny of primates and other mammals. The tree that represents an acceptable phylogeny is shown in Figure 2.3c . This tree was among the 1731 trees that are not significantly different from the maximum likelihood tree. In addition, it was found that trees generated from the same dataset using the neighbour-joining (Saitou and Nei 1987) and maximum parsimony (Fitch 1977) methods were also inconsistent with currently accepted primate phylogeny (data not shown). When the cytochrome c oxidase subunit II gene (Adkins and Honeycutt



(continued over page)

Figure 2.3 - (A) Maximum-likelihood tree found by exhaustive search of tree space (log likelihood score = -3420.10652). **(B)** Consnsus of the maximum-likelihood tree and the 1731 trees not significantly different from it (log likelihood score = -3420.25773). **(C)** Tree most compatible with currently accepted phylogenetic relationships (log likelihood score = -3443.13304). Relative-likelihood support for each branch was determined as described by Jermiin *et al* (1997).

1994) is subjected to the same phylogenetic analysis as described above for cytochrome b, it is similarly unable to resolve the currently accepted phylogeny (data not shown).

Only two groupings of species are clearly resolved by the cytochrome b data. These are the simian primates (*Homo sapiens*, *Colobus guereza* and *Saimiri sciureus*) and the lorisoid strepsirhine primates (*Nycticebus coucang* and *Galago crassicaudatus*). The branching order of the other taxa, including *Tarsius bancanus* and *Lemur catta* cannot be

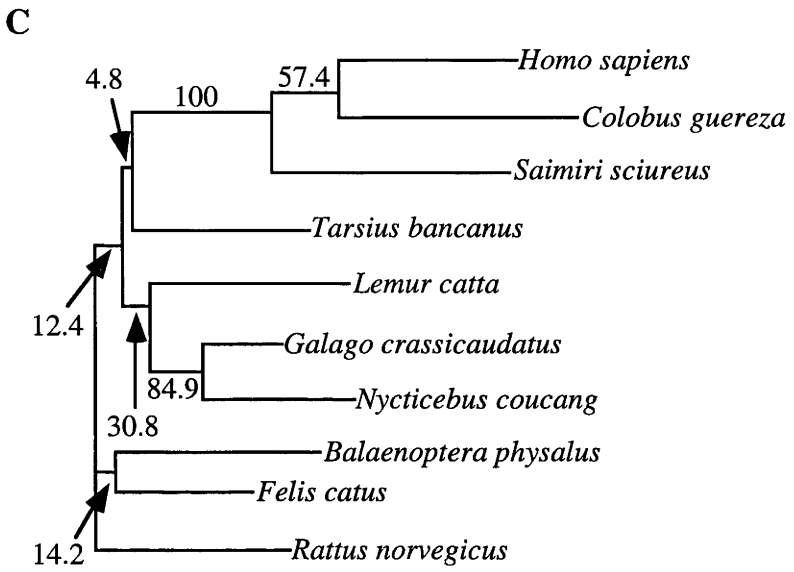


Figure 3.2 - continued.

determined from the data. However, it is evident from Figure 2.3 that the simian primate lineage has undergone more evolutionary change (has longer branch lengths) than all the other lineages, including that of the tarsier, irrespective of which tree is correct. This phenomenon is the same in character as that observed for the cytochrome c oxidase subunit II gene (Adkins and Honeycutt 1994).

Rates of Evolution

Relative-rate tests using a dataset of non-degenerate substitutions is presented in Table 2.3. Using *Rattus norvegicus* as an outgroup, sizable differences were detected in the rate of non-degenerate substitutions between simian primates and other mammals. Focusing on this phenomenon, successive tests were performed using outgroups separated from simian primates by progressively shorter periods of evolution (*Felis catus*, *Lemur catta* and *Tarsius bancanus*). In addition, rate differences among simian primates were analysed using *Saimiri sciureus* as the outgroup. For all comparisons between non-simian mammals and either *Saimiri scuireus* or *Colobus guereza*, substantial increases in evolutionary rate were found for the simian species. *Homo sapiens* also showed an increased evolutionary rate in comparisons with non-simian mammals as demonstrated by large *Z*-scores for comparisons with *Lemur catta* and *Tarsius bancanus*; however these were not

Table 3.3 - Relative rates of substitutions (K) and transversional substitutions (B) per site for the cytochrome b gene from primates and other mammalian species.

Species 1	Species 2	Outgroup	Non-degenerate Subsitutions[a]				Four-fold Degenerate Transversions		
			K_{12} ± SE	K_{13}-K_{23} ± SE	Z-score	K_{01}:K_{02}[b]	B_{12} ± SE	B_{13}-B_{23} ± SE	Z-score
Felis catus	*Homo sapiens*	*Rattus norvegicus*	0.146±0.015	-0.057±0.016	-3.68*	2.290	0.720 ± 0146	-0.123 ± 0.258	-0.479
	Colobus guereza		0.176±0.017	-0.081±0.017	-4.67*	2.686	1.002 ± 0.262	-0.005 ± 0.239	-0.022
	Saimiri scuireus		0.167±0.016	-0.064±0.017	-3.85*	2.246	0.822 ± 0.181	-0.069 ± 0.248	-0.281
	Tarsius bancanus		0.097±0.012	0.005±0.012	0.383	1.431	0.772 ± 0.165	-0.174 ± 0.282	-0.615
	Lemur catta		0.116±0.013	-0.028±0.014	-2.05*	1.303	0.916 ± 0.223	-0.609 ± 0.619	-0.984
	Balaenoptera physalus		0.113±0.013	-0.019±0.013	-1.39	1.116	0.591 ± 0.111	-0.114 ± 0.243	-0.468
Lemur catta	*Homo sapiens*	*Felis catus*	0.147±0.015	-0.030±0.016	-1.93	1.508	0.846 ± 0.193	0.195 ± 0.248	0.789
	Colobus guereza		0.171±0.017	-0.060±0.017	-3.51*	2.075	0.611 ± 0.118	-0.086 ± 0.293	-0.293
	Saimiri scuireus		0.167±0.016	-0.051±0.017	-3.06*	1.887	0.625 ± 0.121	0.093 ± 0.248	0.376
	Tarsius bancanus		0.115±0.013	-0.019±0.014	-1.38	1.390	0.594 ± 0.115	0.144 ± 0.238	0.606
	Nycticebus coucang		0.123±0.014	0.011±0.014	0.821	1.208	0.787 ± 0.171	-0.280 ± 0.408	-0.687
	Galago crassicaudatus		0.101±0.012	0.019±0.012	1.49	1.456	0.596 ± 0.114	0.124 ± 0.240	0.516
Tarsius bancanus	*Homo sapiens*	*Lemur catta*	0.172±0.017	-0.032±0.017	-1.88	1.453	0.815 ± 0.180	-0.251 ± 0.210	-1.198
	Colobus guereza		0.180±0.017	-0.056±0.017	-3.22*	1.909	0.857 ± 0.197	-0.164 ± 0.156	-0.105
	Saimiri scuireus		0.170±0.017	-0.052±0.017	-3.04*	1.875	0.549 ± 0.102	-0.031 ± 0.143	-0.219

(continued)

Table 3.3 - continued.

Saimiri scuireus	*Homo sapiens*	*Tarsius bancanus*	0.147±0.015	0.002±0.016	0.129	1.029	0.527 ± 0.096	-0.266 ± 0.178	-1.491
	Colobus guereza		0.156±0.016	0.010±0.017	0.599	1.140	0.623 ± 0.119	-0.308 ± 0.199	-1.549
Colobus guereza	*Homo sapiens*	*Saimiri scuireus*	0.142±0.015	0.010±0.016	0.612	1.148	0.362 ± 0.64	0.096 ± 0.116	0.827

a - Substitution rates estimated with the two-parameter model (Kimura 1980), and relative rate tests performed using the method of Wu and Li (1985).
b - Average number of non-degenerate substitutions was 713.9, and average number of four-fold degenerate transversions was 71.4.
c - $K_{01} = (K_{13} + K_{12} - K_{23})/2$ and $K_{02} = (K_{12} + K_{23} - K_{13})/2$.
d - $B_{01} = (B_{13} + B_{12} - B_{23})/2$ and $B_{02} = (B_{12} + B_{23} - B_{13})/2$.
* - denotes Z-statistics >1.96 or <-1.96, this being a significance level of a = 0.05.

"significant" at a *Z*-score cut-off of 1.96 (this cut-off would represent an $\alpha = 0.05$ confidence interval if the interspecies comparisons were independent). Comparisons between simian primates showed that little rate heterogeneity exists among these species. From these findings it can be concluded that the cytochrome b gene of simian primates has an elevated evolutionary rate compared to the same gene from the other mammals tested. K_{01}:K_{02} ratios (Table 2.3) show that the magnitude of the rate acceleration is about two-fold when compared to other, non-simian species. These findings are similar to those presented for the cytochrome c oxidase subunit II gene (Adkins and Honeycutt 1994; Ramharack and Deeley 1987).

From these relative rate tests either of two conclusions can be made about the evolution of simian primate cytochrome b. First, the rate acceleration could be the result of an increased mutation rate restricted to the simian lineage, or second, an episode of adaptive evolution could have taken place. To differentiate between these possibilities, additional rate tests were undertaken using a dataset consisting of nucleotide substitutions at four-fold degenerate sites (Table 2.3). However, to do this without obtaining a result that was confounded by the AT/GC compositional heterogeneity at third codon sites (Table 2.2), only transversional substitution rates were compared. Following the same method of outgrouping as before, it could be determined that little rate heterogeneity exists between any of the nucleotide sequences tested. When this lack of rate heterogeneity at four-fold degenerate sites is viewed in relation to the substantial rate heterogeneity observed at non-degenerate sites between simian cytochrome b sequences and those from other mammals, it is evident that the evolutionary forces that have accelerated the rate of substitution for simian cytochrome b have resulted in an increase in non-synonymous changes only. This is strong evidence for positive natural selection having acted on cytochrome b on lineages leading to simian primates and, hence, that simian primate cytochrome b has undergone an episode of adaptive evolution.

The Nature of Amino Acid Replacements in Simian Cytochrome b

Perutz (1983) hypothesised that large modification or refinement of the function of a protein can evolve through changes to only a small number of key residues. This has been

found in a number of cases, such as crocodile haemoglobin (Komiyama *et al* 1995), stomach lysozymes (Kornegay *et al* 1994; Stewart *et al* 1987) and visual pigments (Asenjo *et al* 1994; Yokoyama and Yokoyama 1990; Yokoyama 1995). If simian primate cytochrome b has acquired a new or altered biochemical function of a character similar to that of simian cytochrome c and cytochrome c oxidase subunit II (Osheroff *et al* 1983), then it should be possible to identify the key residue changes that facilitated this. To this end, ancestral cytochrome b sequences were reconstructed for the tree shown in Figure 2.3c using the method of Yang *et al* (1995), and identified amino acid changes that occurred between the most recent common ancestral sequence of all haplorhine primates and the most recent common ancestral sequence of simian primates. Presuming that the altered function of simian cytochrome b arose between these two ancestral sequences, the amino acid changes that occurred along this lineage would include the key changes that caused any functional differentiation. Twenty-six amino acid changes between the ancestral sequences were detected (Figure 2.4).

In an effort to determine which of these amino acid changes were functionally silent or, alternatively, which occurred in important conserved regions of the cytochrome b protein, a variability plot (Figure 2.4) along the length of cytochrome b was constructed using a sliding window approach (after that of Irwin *et al* 1991). For this analysis a larger alignment of nineteen mammalian species (see *Materials and Methods*) was used, including tarsier and strepsirhine primates but excluding simian species, and the absolute number of variable sites in a window twenty residues in length was counted. Using this approach, generally conserved and variable regions of the mammalian cytochrome b protein were identified and these showed correlations with postulated structural regions, such as the Q_i and Q_o redox centres and the trans-membrane domains (Figure 2.4).

Twenty of the twenty-six sites that changed along the ancestral lineage immediately preceding the simian primates are situated in regions of the protein that were classified as non-conserved (defined as containing more than six variant sites per window of twenty sites). The remaining six amino acid changes at positions 39, 42, 57, 60, 67 and 70 are in conserved regions of the protein (defined as containing six or fewer variant sites per win-

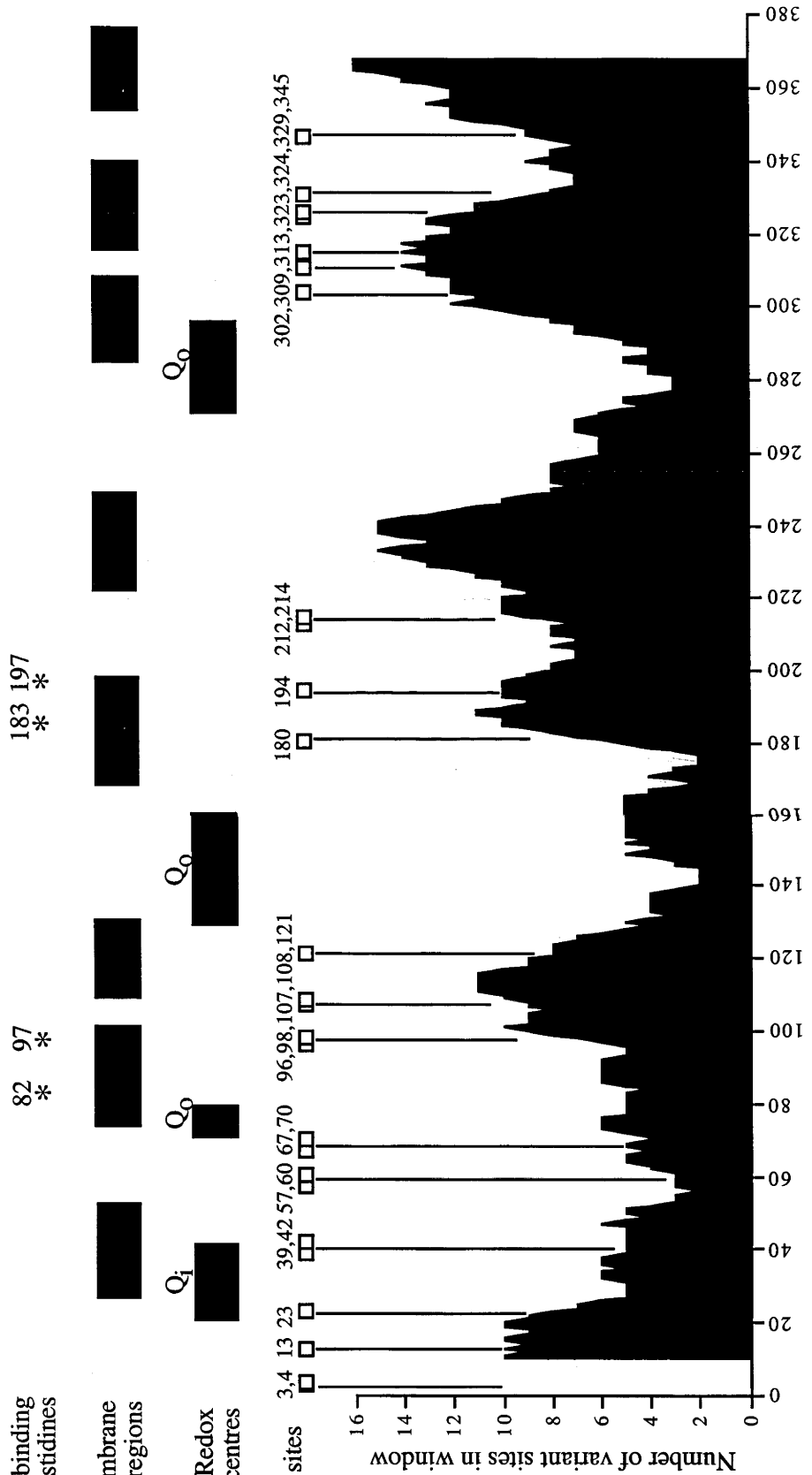


Figure 2.4 - Variablility along the length of an alignment of 19 non-simian primate mammalian cytochrome amino acid sequences, as counted in a sliding window of 20 residues. Also shown are the transmembrane regions from the crystal structure of the beef-heart cytochrome b, key residue changes between ancestral sequences (see text), redox centres and heme-binding histidine residues.

dow of twenty sites). While some of the twenty residue changes that occur in the non-conserved region of the protein could be important for a modified function of simian cytochrome b, it is more likely that the majority of the changes are neutral or compensating changes. The six residue changes that occur in the conserved regions of the protein are more likely to be residues responsible for a change in the function of simian cytochrome b, as they are located in the region proposed to contain the Q_i and part of the Q_o redox sites. At time of writing, the coordinates of the crystal structure of the cytochrome bc_1 complex (Xia *et al* 1997), of which cytochrome b forms a subunit, have not been released, and hence it is not possible to model the structural consequences of the change of this combination of six residues. However, the residues at positions 39 and 42 lie at the end of the Q_i redox site at the N-terminus of the protein, in a region of the protein that has been found to lend resistance to inhibitors in bacterial respiratory chains (Brasseur *et al* 1996). The other residues at positions 57, 60, 67 and 70 are not directly within regions associated with either of the redox sites, but they are on a loop region that is proposed to lie close to the extra-membranous α-helix that forms the central portion of the Q_o redox site. These six amino acid changes that arose in the protein ancestral to extant simian primate cytochrome b may be some of the key residues important in an alteration of cytochrome b function that may have occurred in simian primates but not in other mammals.

Chapter Three

Cytochrome c Oxidase Subunit I of Simian Primates Shows Coordinated Evolution with Other Mitochondrial Electron Transport Chain Proteins.

Abstract

Presented here is a comparative analysis of evolutionary rates between the cytochrome c oxidase subunit I genes of primates and other mammals. Five primate genes were sequenced, and this information combined with existing sequences from four other mammalian orders. The sequences from simian primates show around two-fold increases in non-synonymous substitution rates when compared to other primates and other mammals. The species range and overall magnitude of this rate increase identified for the cytochrome c oxidase subunit I gene is similar to that previously identified for the cytochrome c oxidase subunit II and cytochrome b genes. These three genes encode subunits of the mitochondrial electron transport chain, and the possibility of coordinated adaptive evolution between these genes has been investigated. Here it is shown that possible functional changes to simian cytochrome c oxidase subunit II may have begun an episode coordinated adaptive evolutionary change with several other mitochondrial electron transport chain components, including cytochrome c oxidase subunit I and cytochrome b.

Introduction

The mitochondrial electron transport chain is the enzyme-mediated, biochemical pathway that generates the bulk of cellular adenosine triphosphate (ATP). In simian primates, a number of the protein components of the electron transport chain (ETC) exhibit accelerated evolutionary rates when compared with other mammals, including non-simian primates. The cytochrome b (Chapter 2) and cytochrome c oxidase subunit II (COII) (Adkins and Honeycutt 1994) genes, both encoded in the mitochondrial genome, demonstrate an approximately two-fold increase in non-synonymous substitution rate in simian primates compared to other closely related mammals. These increases in non-synonymous evolutionary rate are strong evidence that these genes have undergone positive selection in simian primates, and may imply an episode of ~~adaptive~~ co-ordinated change has occurred to the ETC in these species. In addition to cytochrome b and COII, nuclear encoded cytochrome c also appears to have undergone a period of rapid evolution on the lineage leading to humans, and this perhaps implies a similar rate increase in other primates (Baba *et al* 1981; Evans and Scarpulla 1988).

Cytochrome c, COII and cytochrome b are closely functionally related. Cytochrome b and COII form integral parts of the final two complexes of the electron transport chain, and cytochrome c acts to transport electrons between them (Hatefi 1985). In addition to the increased evolutionary rates demonstrated by these proteins, it has also been shown that the reaction of simian cytochrome c with the simian cytochrome c oxidase complex, of which COII forms the cytochrome c binding site, is different in protein-protein binding character to that of other mammals (Osheroff *et al* 1983). These facts viewed together have led to the suggestion that these mitochondrial components have co-evolved (Cann *et al* 1984; Chapter 2).

From an analysis of nucleotide substitution rates across mammals in fully sequenced mitochondrial genomes, Janke *et al* (1994) showed that a number of mitochondrial encoded genes violated an assumption of a constant or "clock-like" evolutionary rate. These genes were shown to include, in order of greatest rate violation to least, not only the COII and cytochrome b genes, but also the gene for cytochrome c oxidase subunit I (COI). The recently solved crystal structure for the bovine-heart cytochrome c oxidase complex

(Tsukihara *et al* 1996) shows that COI binds to COII and transduces electrons passed to it from cytochrome c. Potentially, an episode of coordinated adaptive evolution that may involve cytochrome b, cytochrome c and COII could also include COI. This paper presents new COI nucleotide sequences from five primate species along with an investigation of a potential acceleration of evolutionary rate of this gene in simian primates. In addition, analysis of amino acid changes to the COI and COII proteins specific to simian primates is presented in order to identify potential sites of coordinated evolutionary change between these proteins.

Materials and Methods

DNA Sources and Sequencing. Liver samples from *Hapalemur griseus* (grey-gentle lemur) and *Galago senegalensis* (lesser bushbaby) and a lung sample from *Tarsius bancanus* (Bornean tarsier) were obtained from the Duke University Primate Centre. Blood plasma samples from *Ateles geoffroyi* (spider monkey) and *Colobus polykomos* (black-and-white colobus) were donated by the Royal Melbourne Zoological Gardens. DNA was phenol/chloroform extracted (Sambrook *et al* 1989) from either whole plasma samples or homogenates of tissue samples. The COI gene was isolated from each species using the polymerase chain reaction (PCR) with the "universal" oligonucleotide primers shown in Table 3.1. Nucleotide sequences of both strands of each gene were determined directly from the PCR product using dye-terminator cycle sequencing (Perkin-Elmer) with the original PCR primers and additional internal sequencing primers (Table 3.1) on an ABI377 automated sequencer (Applied Biosystems). Additional COI sequences were obtained from published sources: *Balaenoptera physalus* (fin-whale; X61145; Arnason *et al* 1991), *Felis catus* (domestic cat; U20753; Lopez *et al* 1996), *Homo sapiens* (Human; J01415; Anderson *et al* 1981) and *Mus musculus* (mouse; J01420; Bibb *et al* 1981).

Data Analysis. The base composition of each COI gene was determined using the EComposition program of the GCG package (Genetics Computer Group, Wisconsin, USA) available via the Australian National Genomic Information Service (www.angis.org.au).

Base compositional heterogeneity among the COI sequences was assessed using the Distance program (Jermiin *et al* 1998). Pairwise *Z*-scores were calculated between all

Table 3.1 - Oligonucleotide primers used in the amplification and sequencing of COI genes.

Primer Name[a]	Sequence (5'-3')[c]
"Universal" Primers[b]	
H5783	GGCTICTTIGAATTTGCAATTCIA
L7453	TGIGGGTTCGATTCCTTCCTT
Ateles geoffroyi	
H6179	TAGCATTTCCACGAATGAATAA
H6532	CTGACTGACCGTAATCTTAA
H6903	ATGATCTCCTGCAATGCTAT
H7225	TCAGATTACCCCGATGCAT
L7085	AATAGTGGGAAICAGTGAA
L6636	TCCAGGGAGRATAAGGATA
L6283	CTAAGGGTGGGTAAACTGT
Colobus polykomos	
H6171	CCCTGACATAGCATTICC
H6644	TTTTACCAGGCTTTGAATAA
H6881	CACTTCACGGACGCAATAT
H7228	GACTACCCCGACGCTTA
L7100	TTAGAGTATAGCCTGAGAATA
L6804	AAGTGAAATAGGCTCGTGTA
L6324	GGTTAAGTCTACAGAGGCT
Hapalemur griseus	
H6197	ACAACATGAGCTTCTGACT
H6597	TCAACACCTATTCTGATTCTT
H6878	CGACATTACATGGTGGCAA
H7225	TCTGACTATCCAGATGCCT
L7065	GGACGAAICCICCTATAAT
L6700	CCTATATAACCRAATGGTTC
L6287	CCTGCTAGAGGAGGATATA

(continued)

Table 3.1 - continued

Galago senegalensis	
H6271	GGGACCGGATGAACCGT
H6667	CCACATCGTATCCTATTACT
H6973	GTCTTATCAAACTCCTCGTT
L7204	ACGACGAGGCATACCTGA
L6715	TTATTGCCCAGACTATTCCT
L6353	GATACTCCTGCTAGGTGAA
Tarsius bancanus	
H6179	TAGCATTYCCTCGAATAAATAA
H6747	CTTCTTAGGTTTCATTGTCTG
H7080	CTTCGTTCACTGATTCCCA
H7227	CGACTACCCTGACGCATA
L7033	AATAGTGGGAAICAGTGAA
L6636	as above
L6173	TTCGAGGGAATGCTATATCA

a - primer numbers refer to the nucleotide at the 5' most end of the sequence and are numbered with reference to the scheme of Anderson *et al* (1981). H and L refer to the strand to which the primer anneals.
b - universal primers were used for the first sequencing steps from both ends of all genes.
c - I represents deoxyinosine, degenerate bases are represented by standard notation.

pairs of nucleotide sequences at each codon position and the magnitude and range of these graphed to determine which codon sites showed least compositional heterogeneity.

Inferred COI amino acid sequences were aligned using CLUSTALW (Thompson *et al* 1994) and these did not require the insertion of gaps. Hence, COI nucleotide sequences were also aligned without gaps.

The phylogenetic relationship among the COI nucleotide sequences was estimated using the fastDNAml program (Felsenstein 1981; Olsen *et al* 1994) and the TrExML program (Wolf *et al* 1998) was used to perform exhaustive searches of maximum likelihood tree-space. From the TrExML output a list of trees not significantly less likely than the maximum likelihood tree was obtained by performing Kishino-Hasegawa tests (Kishino and Hasegawa 1989) (the K option in the fastDNAml and TrExML programs). A weighted

listing of these trees was derived with the TreeCons program (Jermiin *et al* 1997), and the frequency of occurrence of specific phylogenetic groupings calculated with the Consense utility of Felsenstein's PHYLIP package (version 3.5, University of Washington). These frequencies were used as branch support values on a phylogenetically-constrained tree created with the user-tree option of fastDNAml.

Uniformity of evolutionary rates among nucleotide sequences was tested using the relative rates test of Muse and Weir (1992) and tests of uniformity of synonymous and non-synonymous substitution rates were performed using the method of Muse and Gaut (1994). Additional relative rate tests made using substitution rates estimated by the method of Wu and Li(1985), and *Z*-statistics were calculated using the equation given by Muse and Weir (1992).

Ancestral COI amino acid sequences were estimated with the maximum likelihood method of Yang *et al* (1995) using the codeml program of the PAML package (Version 1.3c, Yang 1997). To assess inter-species variability along the length of the COI and COII protein sequences, additional alignments of these sequences were made. For COI, the sequences aligned previously, except for those of the simians, were translated and additional sequences were included from representative mammalian and marsupial species for which the whole mitochondrial genome has been obtained: *Rattus norvegicus* (Norway rat; X14848; Gadaleta *et al* 1989), *Equus caballus* (horse; X79547; Xu and Arnason 1994), *Equus asinus* (donkey; X97337; Xu *et al* 1996a), *Erinaceous europaeus* (European hedgehog; X88898; Krettek *et al* 1995), *Bos taurus* (cow; V00654; Anderson *et al* 1982), *Rhinoceros unicornis* (Indian rhinoceros; X97336; Xu *et al* 1996b). The alignment of COII sequences was constructed with a similar group of species, also including sequences from *Tarsius bancanus* (Bornean tarsier; L22783), *Galago senegalensis* (bushbaby; M80905), *Lemur catta* (ring-tail lemur; L22780)(Adkins and Honeycutt 1994). Inter-species variability along the length of the COI and COII sequences was assessed from these alignments using a sliding window approach, implemented with the Sliding Window program by TDA with a window size of 20 residues.

Results

Primate COI genes

For this study, five new primate COI nucleotide sequences have been obtained and submitted to the DDBJ/GenBank/EMBL genome database and have sequential accession numbers AB016730-AB016734. These five genes exhibit high levels of similarity among themselves and with other previously published mammalian sequences. Excluding the two simian primate sequences obtained, the new COI nucleotide sequences are 1512 base pairs long and translate to a protein of 513 amino acids, this being the same length as the human sequence. The simian primate sequences obtained for the colobus and spider monkey are unusual as they are longer than the human and other sequences at their 3' end. The colobus sequence is just one codon longer than the other sequences at its 3' end, and the spider monkey sequence appears to incorporate an extra five codons before it reaches a termination codon. As the extra codons and putative termination codon for the spider monkey sequence intrude well into the downstream tRNA gene for serine, ~~and hence~~ this genes is probably post-transcriptionally terminated by the addition of a poly-A tail, as are many mitochondrial genes (Anderson *et al*, 1981; Anderson *et al*, 1982). Fixation of a poly-A tail following thymine at position 1542 of the spider monkey gene forms a termination codon, and termination here makes the spider monkey and human proteins the same size.

Base Composition

Inter-species base compositional heterogeneity has been repeatedly observed in the genes of the mammalian mitochondrial genome. To take account of the inaccuracies that using compositionally heterogenous data can introduce to evolutionary analysis, the COI genes were scrutinised at first, second and third codon positions for possible compositional differences. Table 3.2 shows the frequencies of each nucleotide at each codon position for the COI sequences analysed. From visual inspection of the table it appears that the third codon positions are more compositionally heterogenous than the first and second codon positions. To analyse this difference in base composition in a statistical framework, pairwise comparisons between sequences were conducted to test compositional uniformity of each

Table 3.2 - Base compositions of COI genes at each codon position.

Species	First Codon Position				Second Codon Position				Third Codon Position			
	%A	%T	%G	%C	%A	%T	%G	%C	%A	%T	%G	%C
Human	26.3	22.4	28.8	23.4	19.1	40.5	14.6	25.7	36.0	16.9	5.1	41.8
Colobus	26.7	24.5	27.8	21.0	18.7	41.4	14.6	25.1	35.0	26.7	5.3	32.9
Spider Monkey	27.0	23.5	27.8	21.6	18.9	40.5	14.4	26.1	38.3	27.2	4.7	29.6
Tarsier	26.1	23.3	29.0	21.6	17.9	40.7	14.8	26.5	38.7	27.6	2.9	30.5
Hapalemur	26.3	25.5	30.0	18.3	18.5	40.5	14.8	26.1	40.7	32.1	4.7	22.4
Galago	26.3	24.5	29.2	20.0	17.9	40.9	14.8	26.3	38.1	27.6	6.4	27.6
Whale	27.0	23.7	29.0	20.0	17.9	40.9	15.0	26.1	40.9	21.4	4.7	32.9
Cat	26.7	24.7	29.2	19.5	18.1	40.9	14.8	26.1	34.8	29.2	10.1	25.7
Mouse	25.9	24.1	29.8	20.4	18.3	40.5	14.8	27.7	44.4	27.4	4.1	24.3

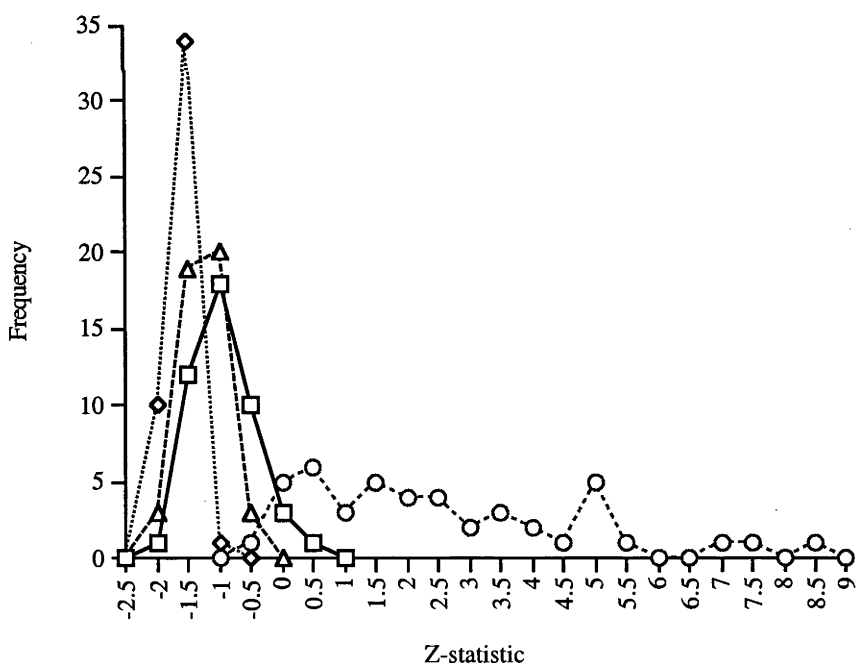


Figure 3.1 - Frequency distribution of COI nucleotide base composition heterogeneity measured as pairwise Z-statistics. First codon position Z-statistic frequencies are represented by squares, second codon positions as diamonds, third codon positions as circles, and first and second positions combined as triangles.

--- ADDENDUM ---
(p45, para 2, replace sentence that begins "In both cases...") -- In both cases maximum likelihood phylogenetic methods did not resolve trees that conformed to currentl accepted primate phylogenies.

pair of sequences (see Jermiin *et al* 1998). Figure 3.1 shows the frequency distribution of *Z*-statistics for comparisons at each codon position. The compositional heterogeneity present at third codon positions, evidenced by the larger range and magnitude of *Z*-statistics, makes these data a poor source of information for analysis of substitution rates. Hence, in following analyses, where it is appropriate, third codon positions have been excluded from the COI nucleotide sequence data-set.

Phylogenetic Reconstruction

Previous analyses of the primate cytochrome c oxidase subunit II (Adkins and Honeycutt 1994) and cytochrome b (Chapter 2) genes have shown that it is difficult to reconstruct reliable phylogenetic trees using the sequences of these genes. see addendum ~~In both cases, trees that reflected currently accepted primate phylogeny could not be derived with maximum likelihood methods.~~ Likewise, the COI sequences could not resolve a realistic phylogeny. Exhaustive search of tree-space (Wolf *et al* 1998) found more than 10000 trees that were not significantly less likely than the maximum-likelihood tree. For the purpose of estimating branch length and evolutionary rates, an assumed phylogeny based on accepted phylogenetic associations derived from other sources (Groves 1989; Irwin *et al* 1991) was applied to the dataset (Figure 3.2). With a log likelihood score of -2999.13337 this tree was among the greater than 10000 trees that were not significantly less likely than the maximum-likelihood tree.

Relative Rates of Evolution

Three series of relative rate tests were conducted to assess the possibility that variant evolutionary rates exist between simian primates and other mammals. Each series of rate tests used a different technique for estimation of evolutionary distance between species.

First, the likelihood-ratio test of Muse and Weir (1992) was applied to first and second codon positions of the COI nucleotide data. Briefly, this method appraised uniformity of substitution rate between pairs of sequences using a maximum-likelihood approach that assumed substitutions at one position in a sequence were independent of sub-

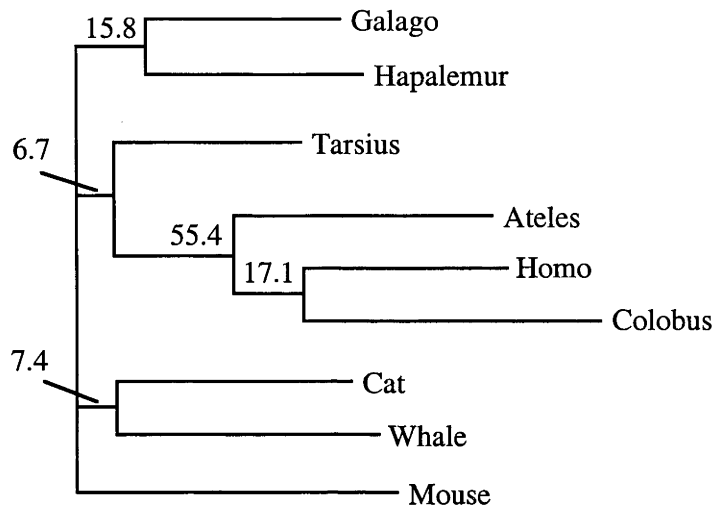


Figure 3.2 - Maximum likelihood tree obtained using a constrained phylogeny. Branch support values are derived from a weighted consensus of phylogenetic associations (type V; Jermiin *et al* 1998) observed from the more than 10000 "most likely" trees generated from exhaustive search of tree-space. Branch support values are very low and reflect the difficulties encountered in generating an acceptable tree from the COI gene sequence data. Note that the overall branch lengths of the simian primates lineages are longer than those of other mammalian lineages.

stitutions at other positions. Multiple log likelihood values were determined for pairs of sequences (and an outgroup) of varying assumptions of equality of substitution rates between the sequences (all substitution rates being equal, transition substitution rates being equal and transversion substitution rates being equal). These values were then compared to a log likelihood value determined with an assumption of different substitution rates. The log likelihood ratios resultant from these comparisons approximate a χ^2-distribution with two degrees of freedom. Applying this method, evolutionary rate variation between the COI genes was investigated. Columns 4-6 of Table 3.3 show the results of tests first looking at relative rates between mammalian orders, and then narrowing in to subsequently test rate differences between primates. While the results of these relative rate tests are not unambiguous, there is a trend of rate differences between simian primates and other mammals. Starting with tests of rate among mammals, no real rate variation, with the exception of that between the cat and colobus, can be seen from the rate tests using the mouse as outgroup. This may be due to the fact that to detect significant rate variation between more highly diverged species it is necessary that larger rate differences be evident (Muse and Weir 1992). When the cat sequence is used as an outgroup, and rate differences

Table 3.3 - Maximum likelihood estimation of rate variation measured by log likelihood ratios[a].

Outgroup	Species 1	Species 2	Non-Codon Based Model[b]			Codon Based Model[c]	
			All Subsitutions	Transitions	Transversion	Synonymous	Non-Synonymous
Mouse	Cat	Whale	2.99	0.12	2.87	2.67	5.26*
		Galago	1.13	0.37	0.76	0.23	2.33
		Tarsier	5.96	5.46*	0.99	1.76	1.72
		Spider Monkey	3.36	0.23	3.09	9.16*	15.83*
		Colobus	9.12*	3.90*	4.97*	7.60*	18.52*
		Human	2.92	0.06	2.83	4.48*	13.13*
Cat	Galago	Hapalemur	4.51	2.25	2.31	0.76	0.15
		Tarsier	1.98	1.86	0.44	0.05	0.42
		Spider Monkey	4.35	4.03*	0.91	0.16	0.26
		Colobus	16.30*	12.52*	3.65*	0.29	5.69*
		Human	8.14*	7.37*	0.73	0.03	0.64
Galago	Tarsier	Spider Monkey	13.61*	3.13	10.38*	3.15	10.97*
		Colobus	20.78*	11.87*	8.66*	2.19	19.18*
		Human	9.64*	0.87	8.74*	10.02*	18.51*
Tarsier	Spider Monkey	Human	0.06	0.00	0.02	0.11	0.00
	Colobus	Human	0.84	0.23	0.58	1.69	0.11
	Colobus	Spider Monkey	0.74	0.69	0.14	2.54	0.21
Spider Monkey	Colobus	Human	11.47*	6.63*	4.64*	1.81	10.92*

a - Asterisks denote significance at a confidence level of α=0.05 assuming that the data resembles a χ^2 distrubution with two degrees of freedom.

b - Muse and Weir (1992).

c - Muse and Gaut (1994).

assessed between the galago and other primates, smaller evolutionary distances are being appraised and significant rate variation can be seen for comparisons with simian primates (although the spider monkey sequence does not show an overall rate difference). Comparisons between simians and the galago, with a cat outgroup, show that the source of this rate difference is largely due to transitional substitutions. However, the rate tests between the tarsier and the simians (with a galago outgroup) show the opposite, and the source of the rate difference is transversion substitutions. The rate tests conducted between the simian primates show little rate variation although the colobus sequence may have evolved at a different rate to the human and spider monkey sequences. From these tests it is only possible to identify rate differences between species, not which species has the greater rate. However, the longer branch lengths evident for the simian species on the phylogenetically constrained tree shown in Figure 3.2 would imply that the simians have had higher evolutionary rates than other mammals. This is investigated further below.

A second series of relative rate tests was conducted using the method of Muse and Gaut (1994). This method uses a codon0based model to assess relative substitution rates, and takes into account substitutional dependencies that nucleotides have within a codon (ie. some substitutions will be silent and hence exposed to different evolutionary forces than other substitutions that cause amino acid changes in the expressed protein). Like the previous method, this method calculated likelihood values with different rate assumptions (synonymous substitution rates equal, non-synonymous substitution rates equal compared with different synonymous and non-synonymous substitution rates) and employed a likelihood ratio test as to assess rate differences. Due to the codon-oriented nature of these tests, use of all codon positions was required. This requirement would appear undesirable in the case of the COI genes where the third codon site demonstrated base compositional heterogeneity, however Muse and Gaut (1994) claim that their model adequately takes account of such biases. The last two columns of Table 3.3 show the results obtained with this method employing the same rate test strategy as applied before. The results indicate that rate differences are apparent between the simian primate sequences and those of other mammals. Rate tests among mammals with the mouse used as an outgroup show signifi-

cant log likelihood ratios between the cat and simian sequences. These comparisons show both synonymous and non-synonymous substitution rates are increased, but the likelihood ratio values for non-synonymous substitutions are two to three times larger than those for synonymous substitutions. Rate tests between the galago and other primates, using the cat sequence as an outgroup, do not show the same pattern of rate variation between simians and other mammals, with the exception of the comparison with the colobus sequence. This result viewed with the results of separate sets of rate tests above (and see below) suggest that the galago sequence may have an altered rate that obscures the rate differences otherwise observed for the simian sequences. Conversely, rate tests between simian primates and the tarsier, with a galago outgroup, show substitution rate differences between these species, with significant likelihood ratio values overall and at non-synonymous sites. Rate tests among simian primates show little rate variation exists between these species, although the rate tests viewed together suggest that the colobus sequence may exhibit a slightly higher non-synonymous substitution rate than the human and spider monkey sequences.

Given the large number of relative rate tests summarised by log-likelihood ratios made here, it is important to note that multiple comparisons of this kind suffer from type-I error (erroneous rejection of the null hypothesis). To avoid this, Bonferroni corrections (Rice 1989) need to be made to determine the "table-wide significance" of the results obtained. After correction of the data shown in Table 3.3 only four tests showed significant rate difference between species (data not shown, but see the diskette appendix). Using the nucleotide model there was a rate difference between the cat and spider monkey (significant at α=0.15), and from the codon model tests there were non-synonymous rate differences between the cat and the colobus, the spider monkey and the human (significant at α=0.01, 0.005 and 0.05, respectively). While this result casts doubt over the inferences made above, it is very important to note that in cases where data is not independent, as in a relative rate test, Bonferroni corrections are overly conservative. The reality of the rate differences between simian primates and other mammals lies somewhere between the results shown in Table 3.3 and the results of the Bonferroni correction.

The third method used to investigate potential substitution rate variation between simian primates and other mammals was that of Wu and Li (1985). This method treats substitutions differently depending on their context within a codon and whether they occur at non-degenerate, two-fold degenerate or four-fold degenerate sites. Unfortunately, the codon orientation of this method requires the use of third codon positions, but the utility of the method for measuring evolutionary distance in degeneracy classes allows a measure of synonymous (four-fold degenerate sites) and non-synonymous (non-degenerate sites) substitution rates. Substitutions at two-fold degenerate sites are also be classified as synonymous if they are transversions or non-synonymous if they are transitions, however, there are exceptions to this rule and although these are minor they complicate the analysis and hence have not been included. The same set of rate tests as applied previously were repeated using this method, and the results of these are shown in Table 3.4. Here, four-fold degenerate transitions are excluded from the analysis to avoid possible inaccuracy introduced by third codon position AT/GC bias. *Z*-statistics were calculated for each rate test (data not shown) to statistically appraise rate differences between species, however none were significant. This lack of significance is more a result of the large variances associated with each of the rate estimates shown in Table 3.4, rather than an indication of a lack of rate variation. Hence, due to the uninformative nature of the *Z*-statistics, rate differences between species were appraised through comparison of K_{01} and K_{02} distances. Rate tests conducted between the cat sequence and other mammals, using a mouse outgroup, show at non-degenerate sites that the cat sequence has the highest rate among non-simian primate sequences, but that all simian have a higher rate than the cat. The opposite of this is apparent at four-fold degenerate sites however. Focusing on the primates, rate tests between the galago and the hapalemur, and the galago and the tarsier show that the galago has the higher rate at both non-degenerate and four-fold degenerate sites. Comparisons between the galago and simian primates, however, show that the simians have a higher evolutionary rate at both non-degenerate and four-fold degenerate sites (although the rate difference is smaller at four-fold degenerate sites). Tests between the tarsier sequence and the simian sequences show the simians to have higher evolutionary

Table 3.4 - Rate variation between COI sequences assessed per 100 sites using distances calculated by the method of Wu and Li (1985).[a]

Outgroup	Species	Species	Non-degenerate Subsitutions				Four-fold Degenerate Transversions			
(3)	(1)	(2)	$K_{12} \pm$ SE	K_{13}-$K_{23} \pm$ SE	K_{01}:K_{02}[a]	K_{02}:K_{01}[a]	$B_{12} \pm$ SE	B_{13}-$B_{23} \pm$ SE	B_{01}:B_{02}[a]	B_{02}:B_{01}[a]
Mouse	Cat	Whale	3.36 ±0.6	-0.37±11.7	1.46	0.684	93.8 ±19.7	-65.8±102.8	1.11	0.897
		Galago	3.24 ±0.6	-0.29±11.6	1.56	0.643	113.3 ±29.3	3.3±66.9	0.812	1.23
		Tarsier	3.04 ±0.6	0.97±11.2	2.14	0.467	106.8 ±22.5	6.2±65.8	0.925	1.08
		Spider Monkey	5.38 ±0.8	-1.75±12.1	0.675	1.48	94.4 ±19.9	-52.5±93.1	1.12	0.895
		Colobus	6.92 ±0.9	-3.07±12.5	0.461	2.17	92.8 ±19.4	-57.4±96.7	1.14	0.881
		Human	5.29 ±0.8	-1.65±12.1	0.694	1.44	100.4 ±22.4	-57.1±96.4	1.03	0.968
Cat	Galago	Hapalemur	2.83 ±0.5	0.35±10.6	1.33	0.752	101.7 ±23.2	20.2±68.9	1.29	0.777
		Tarsier	2.77 ±0.5	0.20±10.7	1.36	0.737	56.6 ±8.9	6.6±70.8	3.26	0.306
		Spider Monkey	5.94 ±0.8	-2.14±11.6	0.332	3.01	151.2 ±62.5	18.9±69.9	0.427	2.34
		Colobus	7.10 ±0.8	-3.67±12.0	0.283	3.53	123.2 ±35.9	20.5±69.2	0.809	1.24
		Human	5.36 ±0.8	-2.04±11.5	0.427	2.34	-[b]	-[b]	-[b]	-[b]
Galago	Tarsier	Spider Monkey	4.76±0.7	-3.16±11.5	0.500	2.00	91.5±18.6	-94.6±83.5	0.626	1.60
		Colobus	6.45±0.8	-4.32±11.7	0.318	3.14	134.8±45.2	-66.6±66.8	0.224	4.47
		Human	5.10±0.7	-2.58±11.4	0.396	2.53	92.1±18.8	-[b]	-[b]	-[b]

(continued)

Table 3.4 - continued.

Tarsier	Spider Monkey	Human	4.66 ±0.7	-0.34±12.0	1.04	0.962	44.0 ±6.6	-0.68±58.8	3.13	0.319
	Colobus	Human	6.09 ±0.8	1.35±12.5	1.15	0.868	30.8 ±4.8	42.7±74.1	16.54	0.060
	Colobus	Spider Monkey	7.66 ±0.9	1.69±12.4	0.595	1.68	63.2 ±10.4	43.4±76.9	8.24	0.121
Spider Monkey	Colobus	Human	6.09 ±0.8	3.00±12.7	2.02	0.496	30.8 ±4.8	19.2±39.6	6.55	0.153

a - $K_{01} = (K_{13} + K_{12} - K_{23})/2$, $K_{02} = (K_{12} + K_{23} - K_{13})/2$, $B_{01} = (B_{13} + B_{12} - B_{23})/2$, and $B_{02} = (B_{12} + B_{23} - B_{13})/2$

b - the pairwise comparison of the galago and human sequences at four-fold degenerate transversion sites resulted in a division by zero error in the calculation of the Q value (see Wu and Li 1985).

rates, once again at both non-degenerate and four-fold degenerate sites. Tests amongst the simian primates themselves are not entirely consistant, but imply little rate difference exists between these species at non-degenerate sites, but extreme rate difference is found at four-fold degenerate sites. Taken together these results show that the simian primates in general have a higher evolutionary rate than other mammals. The elevation in non-degenerate substitution rate appears to be more predominant than the four-fold degenerate substitution rate (at least in comparisons between simians and non-primate mammals). This is evidence that either non-degenerate change to the COI gene has been specifically selected, or that four-fold degenerate sites have become saturated. Figure 3.3 shows overall non-degenerate evolutionary distance (K_0) plotted against four-fold degenerate distance (K_4) for comparisons between all COI sequences. While there is not a clear trend, this shows that K_0 values are roughly proportionate to K_4 values, and saturation at four-fold degenerate sites is not apparent, although this cannot be ruled out. From this information, it is probable that (at least between non-primate mammals and simian primates) there has been an increase in the non-degenerate substitution rate for simian primate COI sequences, and that this has not been accompanied by an increase in the four-fold degenerate substitution rate.

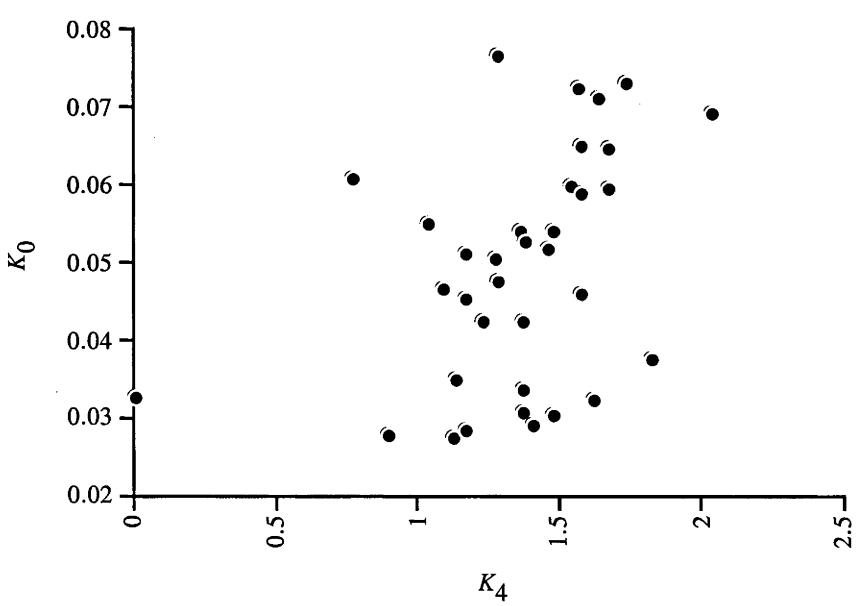


Figure 3.3 - Non-degenerate evolutionary distance (K_0) plotted against four-fold degenerate distance (K_4) calculated using the method of Li *et al* (1985) for comparisons of all COI genes.

--- ADDENDUM ---

(p54, para 1, second last sentence) -- As noted in chapter 2, even though the pattern of non-synonymous rate increases in the COI genes for simian primates strongly suggest some sort of adaptive change, it is not possible to wholly rule out a relaxation of functional constraint due to these genes not displaying K_a/K_S ratios greater than one (see Table 3.4).

When the results of the relative rate tests using the three different methods are taken together, an increase in the substitution rate of simian primates is evident compared to other mammals, including non-simian primates. Furthermore, it is possible to determine the source of this increase and show that this has largely been due to an increase in the rate of non-synonymous substitution. Such a pattern of evolutionary rate increases at non-synonymous sites is evidence for possible adaptive evolution having occurred to the COI gene of simian primates. Most interestingly, this pattern evident for COI is very similar to that demonstrated by the functionally related COII and cytochrome b ETC proteins.

Potential Adaptive Changes of Simian COI and COII Proteins

If the accelerated rate of the simian primate COI gene can be attributed to adaptive evolution, then it would be of interest to identify the particular substitutions that may have been involved. As the increased rate of evolution has been identified in simian primates, an adaptive event that would affect all of these species would have occurred after their divergence from tarsiers, but before their divergence from their last common ancestor. Identifying amino acid changes that occurred along this lineage may identify the changes that could have caused a functional change or modification to the COI protein (and therefore caused this gene to be selected). To investigate this, ancestral amino acid sequences for the COI protein were estimated using the maximum likelihood method of Yang *et al* (1995). From these ancestral sequences it was possible to estimate the amino acid changes that occurred on the lineage between the last common ancestor of Haplorhine primates and the last common ancestor of simian primates. Thirteen changes were detected along this lineage (see Figure 3.4a), and presumably any phenomenon that affected all simian primates would have been due to all or a subset of these changes. To determine the functional significance of these key residues may have had, the thirteen amino acids were assessed with regard to what is known of the functional role of subunit I in the cytochrome c oxidase complex. A crystal structure of cytochrome c oxidase has been obtained from both cows (Tsukihara *et al* 1996; Tsukihara *et al* 1995) and the bacteria, *Paracoccus denitrificans* (Iwata *et al* 1995). In conjunction with this, a sliding window approach was

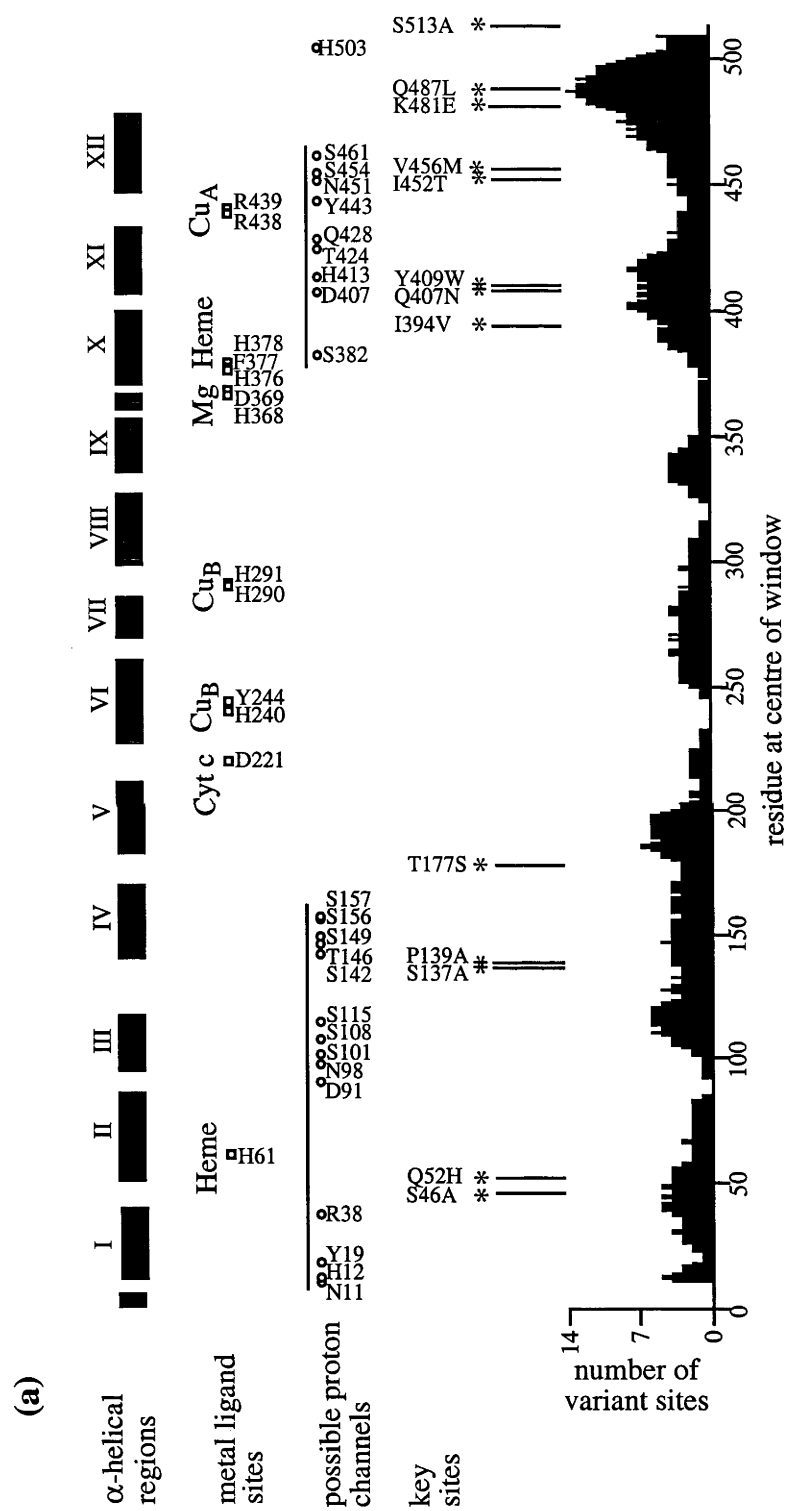


Figure 3.4 - Variability along the length of the mammalian (**a**) COI (above) and (**b**) COII (next page) proteins. Positions of a-helical transmembrane domains determined from the crystal structure (Tsukihara *et al* 1996) are shown by Roman numerals. Also shown are the positions of metal ligands (Tsukihara *et al* 1995), cytochrome c docking sites (Witt *et al* 1998a; Witt *et al* 1998b) and possible proton channels (Tsukihara *et al* 1996). Substitutional changes on the lineage between the last common ancestor of Haplorhine primates and the last common ancestor of simian primates determined from reconstructed ancestral sequences are marked with an asterisk.

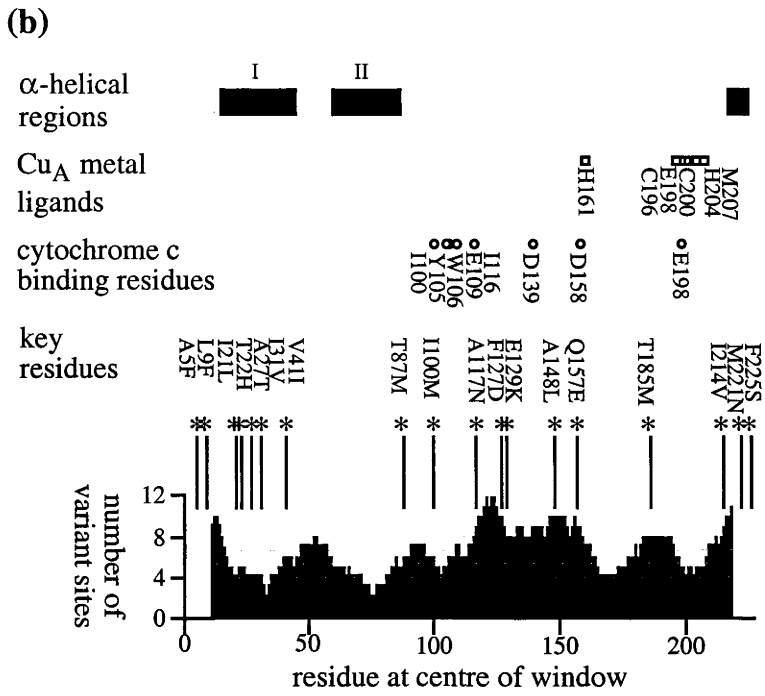


Figure 3.4 - continued.

used to assess the inter-species variability of the COI protein in a representative set of mammals (see *Materials and Methods*). Superimposition of the variability of the mammalian COI protein with structural information obtained by the crystallographic studies (Figure 3.4a) allows indirect estimation of the importance of that any of the key residue changes to COI may have had to the function of the simian cytochrome c oxidase complex. The COI protein was found to be highly conserved between mammalian species, as compared to other mammalian mitochondrial proteins, such as cytochrome b (Chapter 2, Figure 2.4) and COII (Figure 3.4b, see below). COI forms the core of the cytochrome c oxidase complex, contains the enzymes electron transport mechanism, putative proton pumping channels and the dioxygen reduction site, and it is not unexpected that the protein is well conserved. The location of the key residue changes shown in Figure 3.4a indicate that they all fall into the more variable regions of the subunit, and avoid important structural residues. However, this does not mean the changes are important, and a number fall in and around clusters of important residues. Most notably, the changes at positions 137, 139, 407, 409, 456 and 452 (see Figure 3.4a) lie in regions of the protein that are postulated to be important in proton pumping. Hence, while the highly conserved struc-

tural residues of COI are not altered in the simian proteins, a number of residue changes specific to simians may have some indirect effect on the function and/or efficiency of these regions. Detailed structural modelling of these changes will provide more information on their potential importance.

Early structural predictions (Capaldi *et al* 1983) and the more recent crystal structure of cytochrome c oxidase show that the COI and COII proteins come in to direct contact in a large number of places in the cytochrome c oxidase holoenzyme. In particular, the two transmembrane helices of COII come into contact with transmembrane helices VIII and IX of COI, and the carboxy-terminal extra-membrane loop of COII which contains the cytochrome c binding site overlies the matrix-side extra-membrane loops of COI. To investigate if potentially important evolutionary changes to COI may have occurred to these regions of the protein subunit, the variability of COII was assessed in regard to structural information in the same way as was done for COI (Figure 3.4b). In addition, ancestral sequences were estimated for COII and potentially important residue changes also predicted in the same way as was done for COI. From this information, it was possible to get a preliminary indication of whether any of the key changes in simian COI or COII are potentially involved in interactions between these proteins. From comparison of Figures 3.4a and b, it would seem unlikely that there has been any adaptive co-evolutionary change in the transmembrane helical regions of the subunits, since none of the key changes identified in COI fall in either of the helices that interact with COII. There are a number of changes that have occurred in the transmembrane helix I of COII, but as observed by Adkins and Honeycutt (1994), these are conservative changes which preserve the hydrophobic nature of the helix. Figure 3.4b shows many changes have occurred in the carboxyl-terminus extra-membrane region of COII, and these may be involved in interactions with the exposed extra-membrane loop regions of COI (between helices V and VII, and VIII and IX). However, Figure 3.4a shows that none of residue the changes identified occur in these loop regions of COI. Hence, while COII may have changed to interact with COI, COI has not changed to interact with COII.

From Figure 3.4b, it is apparent that the COII protein is not quite as conserved as the COI protein. While most of the residue changes identified in COII lie in less-conserved regions of the protein, so do a number of conserved structural residues identified from the crystal structure. Simian COII demonstrates, perhaps most importantly, a change from isoleucine to methionine at position 100, a residue inferred from the *Paracoccus* crystal structure as important for cytochrome c binding (Witt *et al* 1998a). Previous studies have shown modified binding characteristics of simian cytochrome c to the simian cytochrome c oxidase complex, when compared to other mammals (Osheroff *et al* 1983). Potentially an I to M change at position 100 of simian COII may have had some role in this. This change was the only change to actually coincide with a postulated important structural residue of COII, but as was the case for COI, many other changes specific to simian COII lie close to important functional regions of the protein, particularly those shown to be important in cytochrome c binding. The simian-specific residue changes identified in COII suggest that the evolutionary rate acceleration observed for COII may be related to the altered interaction of this protein with cytochrome c.

Discussion

Previous investigations of accelerated evolutionary rates in primate mitochondrial genes have shown that the cytochrome b (Chapter 2) and COII (Adkins and Honeycutt 1994) genes have increased rates of non-synonymous substitution in simian primates. The analysis conducted here presents further evidence that these rate increases may be part of a multi-enzyme evolutionary phenomenon, and has shown that COI sequences from simian primates have undergone an evolutionary rate acceleration in the same species range as both the cytochrome b and COII genes. Each of these three mitochondrial genes show a rate acceleration in simian primates of approximately the same magnitude, being about a two- to three-fold increase in non-synonymous substitution rates compared to that of other closely related mammals. These increases in evolutionary rate for simian primates also appear not be uniform among the simian primates, and for each gene the Old World monkeys (baboons, macaques, colobine monkeys) show the greatest increase in substitution rate.

The three above-mentioned mitochondrial genes also share the quality that they were all refractory to successful phylogenetic reconstruction, indicating that inter-species rate heterogeneity can perhaps be as confounding to obtaining a correct phylogeny as base compositional heterogeneity. However, the difficulties encountered here could also be due to single genes being used, and more sequence information would allow a true phylogeny to be obtained. Studies that use these simian genes to conduct phylogenetic analyses need to view their results with caution. In particular, it is unfortunate that the cytochrome b gene is one of the most popular genes used for small and large scale phylogenetic work. Results obtained using the COI, COII and cytochrome b genes from simian primates should be viewed carefully in the light of the extreme inter-species rate heterogeneity demonstrated for these genes from simians and other mammals.

From the pattern of evolutionary rate increases shown here and from previous studies (Adkins and Honeycutt 1994; Chapter 2), it has become possible to predict the range of the coordinated adaptive evolutionary change that may have occurred between these components of the electron transport chain. So far it appears that both complex III (the cytochrome bc_1 complex, which includes cytochrome b as a major functional subunit) and complex IV (cytochrome c oxidase, which contains both COI and COII) of the electron transport chain are involved, along which cytochrome c as the functional link between the two complexes (see Box 1.2 and Figures 1.2 and 1.3). In complex III, cytochrome b forms a large membrane spanning core protein around which other subunits of the complex cluster (Xia *et al* 1997). However, it is evident from the crystal structure of complex III that cytochrome b itself does not come directly into contact with cytochrome c. Hence, if cytochrome b is part of a coordinated episode of adaptive evolution involving cytochrome c and subunits I and II of cytochrome c oxidase, then other non-mitochondrial encoded subunits of complex III must also be involved. Complex III subunits that are involved in passing electrons from cytochrome b to cytochrome c would be prime candidates, and in particular this includes the Rieske iron-sulfur protein and cytochrome c_1.

As subunit III of cytochrome c oxidase appears not to show an accelerated evolutionary rate (Janke *et al* 1994; Chapter 4), and also is not ascribed with a functional role in interacting with cytochrome c (Tsukihara *et al* 1996), it is probable that this protein has

not been part of coordinated evolutionary changes that may have taken place within the cytochrome c oxidase complex. In addition, other nuclear encoded subunits of the cytochrome c oxidase complex appear unlikely to be involved in co-evolution with COI and COII, as they do not come into functional contact with cytochrome c, nor do they have significant roles in electron transduction (Tsukihara *et al* 1996). Hence, the scope of coordinated evolution that may have occurred within the cytochrome c oxidase complex appears restricted to COI and COII. From defining the scope of this proposed episode of coordinated adaptive evolution, it seems likely that the trigger of co-evolution was a change in the interaction between cytochrome c and COII. Biochemical data has shown that this interaction for simian primates is specific to simian primates (Osheroff *et al* 1983), and hence this may be where a period of co-adaptive change began. Analysis of the key amino acid changes on the lineage leading to simian COII shows that a number of these changes could have had an effect on the function of the cytochrome c oxidase complex in regard to its interaction with cytochrome c. However, a similar analysis of COI on the same lineage showed that it appeared to have largely undergone conservative changes. This information viewed as a whole tends to suggest an evolutionary change has occurred in the cytochrome c oxidase complex on the immediate lineage leading to simian primates, and COII with its interaction with cytochrome c may have been the site of the original change. Other proteins such as COI and cytochrome b may have had to change subsequently to compliment changes in both COII and cytochrome c to maintain functionality. If this were the case, this would explain why amino acid changes to both COI and cytochrome b on the lineage leading to simian primates are not highly likely to have changed the function of the proteins, and why the rate of evolution of these proteins remains accelerated after the divergence of the simian primate species from their last common ancestor.

Chapter Four

A Survey of Mitochondrial Protein-Coding Genes Which Have Increased Non-Synonymous Evolutionary Rates in Primates.

Abstract

The constancy of evolutionary rates for each of the protein coding genes of the mammalian mitochondrial genome was investigated with the aim of mapping the extent to which a proposed episode of adaptive evolution has occurred in the mitochondria of simian primates. Previous studies have shown an increase in the non-synonymous substitution rate in a number of mitochondrial genes on the lineage immediately preceding the most recent common ancestor of simian primates. By comparing evolutionary rates of mitochondrial genes between ape and other mammalian species, an assessment was made as to which simian genes may have been part of a coordinated adaptive episode. The cytochrome c oxidase subunit II and cytochrome b genes showed distinct rate increases in apes, while the cytochrome c oxidase subunit I, NADH dehydrogenase subunits 1 and 5 and ATP synthase subunit 8 genes showed smaller increases. Other mitochondrial genes were found to have rates either similar (NADH dehydrogenase subunits 3 and 6, and ATP synthase subunit 6) or less than (NADH dehydrogenase subunits 2, 4 and 4L) other mammals, implying that the protein-coding genes of the mitochondrial genome of mammals have evolved as distinct loci. Lineage specific estimates non-synonymous and synonymous substitution rates for the cytochrome c oxidase subunit I and II, and cytochrome b genes did not convincingly identify the lineage where an episode of positive selection may have occurred. Given the increase in non-synonymous substitution shown previously for these genes, this finding may mean that this method is too conservative for detecting all but the most pronounced cases of selection.

Introduction

Several mitochondrial protein-coding genes have been shown to have evolved more rapidly in simian primates than in other mammals. In particular, the cytochrome c oxidase subunits I and II (COI and COII) and cytochrome b (COB) genes show evolutionary rate increases of up to three times on lineages leading to simian primates compared to those of other mammals (Adkins and Honeycutt 1994; Chapters 2 and 3). These rate differences appear restricted to non-synonymous substitutions, and synonymous substitution rates appear to have remained unchanged. While saturation of synonymous substitutions may be partly responsible for this pattern, this finding overall suggests that an episode of coordinated adaptive evolution may have taken place between these genes in simian primates.

The full scope of a potential episode of adaptive co-evolution in the mitochondrial genome of simian primates has not been assessed, however it has been shown that nuclear-encoded cytochrome c gene also exhibits a substantial increase in amino acid substitution rate in a similar group of primates as that found for the mitochondrial genes (Baba *et al* 1981; Evans and Scarpulla 1988). The proteins encoded by the COI, COII and COB genes are all components of the multi-enzyme mitochondrial electron transport chain and are key subunits of the final two complexes - complexes III (cytochrome c oxidoreductase) and IV (cytochrome c oxidase)(Hatefi 1985). Biochemically, complexes III and IV are linked by cytochrome c, which functions in shuttling electrons from complexes III to IV. In performing this function cytochrome c must bind both complexes, and it has been shown *in vitro* that the nature of the interaction of cytochrome c with complex IV in simian primates is different to the interaction in a range of other mammals (Osheroff *et al.* 1983). Potentially, adaptive co-evolution among the mitochondrial COI, COII and COB genes and nuclear-encoded cytochrome c may be part of a larger phenomenon that involves a range of both mitochondrial- and nuclear-encoded genes and other complexes of the mitochondrial electron transport chain.

The increase in evolutionary rate in these three mitochondrial genes appears to have begun on the lineage leading to the most recent common ancestor of simian primates, following the simian-tarsier divergence (Adkins and Honeycutt 1994; Chapters 2 and 3). If the co-adaptive phenomenon observed for these simian genes extends to other mito-

chondrial genes, existing sequence information from complete mitochondrial genomes is adequate to gain a preliminary indication of which genes show an evolutionary rate acceleration in primates. Through comparing the substitution rates of mitochondrial genes of apes with other mammals, the relative rates of evolution of these genes can be inferred. While the exact lineage along which any rate differences may have occurred cannot be identified, such an analysis will give an indication of which mitochondrial genes may have accelerated rates in primates. This information will indicate which genes in addition to the simian COI, COII and COB genes may also be involved in an episode of coordinated adaptive evolution in simian primates.

Materials and Methods

Data sources. Nucleotide sequences for each of the protein coding genes of the mitochondrial genome were extracted from whole genome sequences published previously; human (*Homo sapiens*, J01415; Anderson *et al* 1981), gorilla (*Gorilla gorilla*, D38114; Horai *et al* 1994), domestic cat (*Felis cattus*, U20753; Lopez *et al* 1996), whale (*Balaenoptera physalus*, X61145; Arnason *et al* 1991), horse (*Equus callabus*, X79547; Xu and Arnason 1994), and mouse (*Mus musculus*, J01420; Bibb *et al* 1981). Additional primate nucleotide sequences for the COB, COI and COII genes used for estimating synonymous and non-synonymous substitutions on internal branches were obtained from the following sources; COI (Chapter 3), COII (Adkins and Honeycutt 1994), COB (Chapter 2; Irwin *et al* 1991).

Analysis. The inferred amino acid sequence of each of the mitochondrial genes was aligned using CLUSTALW (Thompson *et al* 1994), and gaps introduced in this alignment were then manually transferred to an alignment of nucleotide sequences . By aligning the nucleotide sequences in this way it was possible to ensure that gaps did not alter the reading-frame. Small variation in gene length at the end of the sequences of a number of genes was encountered, and so as not to unduly bias the rate tests these were truncated to the length of the shortest sequence. In doing this, at most, four codons were discarded from any one sequence, and all stop codons were removed.

Base frequencies for each gene were determined using the codeml function of the

PAML package (Yang 1997), and base heterogeneity at each codon position was assessed using the Distance program (Jermiin *et al* 1998).

Nucleotide- and codon-based maximum-likelihood estimates of rate differences were calculated using the methods of Muse and Weir (1992) and Muse and Gaut (1994), respectively and implemented using the codrates program by S. Muse (available by anonymous FTP from bio.indiana.edu).

Maximum-likelihood estimates of synonymous and non-synonymous substitutions and likelihood-ratio tests of uniformity of dN/dS ratios on internal branches of gene phylogenies were conducted by the method of Yang (1998), and implemented by version 1.3c of the PAML package (Yang 1997).

Results

Relative rate tests

Evolutionary rate differences have been appraised between ape and other mammalian species to determine whether additional mitochondrial protein coding genes show an elevated rate of non-synonymous substitution similar to that shown previously for the simian COB, COI and COII genes. While comparisons between apes and other mammals will not show the exact lineage on which an evolutionary rate acceleration may have occurred (such as was done in Chapters 2 and 3), it will identify other genes that display higher evolutionary rates in primates, which would be worthy of further investigation when more sequence information becomes available. The maximum-likelihood relative rates tests of both Muse and Weir (1992) and Muse and Gaut (1994) have been used to test hypotheses of evolutionary rate uniformity between representative ape species and other mammals.

Exhaustive application of the relative rate tests of Muse and Weir (1992) and Muse and Gaut (1994) to the dataset of thirteen different mitochondrial genes from two apes (human and gorilla) and four other mammals (cat, whale, horse and mouse) could potentially lead to the computation of over a thousand test statistics. Such analysis leads to a statistical problem due to multiple comparisons, and can lead to type-I error (erroneous rejection of the null hypothesis). In addition, the non-independence of each gene sequence means that Bonferroni corrections (Rice 1989) will be conservative (type-II error,

erroneous acceptance of the null hypothesis). Hence, to counter this the number of relative rate tests conducted using these maximum-likelihood methods have been restricted by first comparing non-synonymous to synonymous (dN/dS) ratios between ape species and other mammals, and considering test statistics (log-likelihood ratios) only of genes that have universally higher ratios in apes. dN/dS ratios were calculated using the method of Muse and Gaut (1994), assuming a model where neither non-synonymous nor synonymous substitution rates were constrained to uniformity between lineages. Ratios were compared between the ape lineage and the other mammalian in-group lineage, and a positive result counted if the ape lineage had a higher ratio. Six comparisons were possible between the ape and other mammal in-group species (human-cat, gorilla-cat, human-whale, gorilla-whale, human-horse, gorilla-horse), and each gene appraised according to the number of positive results returned. Table 4.1 shows the list of mitochondrial genes ordered according to the number of positive results they returned. Simple means of the dN/dS ratio on the ape lineage are also shown for reference. If a given gene had six positive results then this was a remarkable correlation, and may mean that this gene has a higher dN/dS ratio. For genes which showed six or five postive results, log-likelihood ratio test statistics were calculated for the human-cat comparison. By calculating just a small number of test statistics in this manner, major type-I errors have been circumvented, although potentially one significant value may be erroneous. Table 4.1 shows log-likelihood ratios calculated using both the codon based method of Muse and Gaut (1994) and the nucleotide method of Muse and Weir (1992). The codon model test statistics have been calculated on a dataset consisting of all codon positions, while the nucleotide model test statistics have been calculated from just first and second codon positions. By omitting third codon positions it is possible to avoid most of the base compositional heterogeneity that was found to be present for each gene between species (data not shown, but see the diskette appendix). The test statistics presented in table 4.1 show that four to potentially six mitochondrial genes may have increased rates of non-synonymous substitution on ape lineages. Using the codon model the ND5, ATP8, COII and COB genes show significant non-synonymous rate differences, which can be inferred as being rate increases. These

Table 4.1 - Maximum-likelihood relative rate tests.[a].

Gene	Positives	Primate Lineage Substitution Rate			Codon Model		Nucleotide Model
		Non-Synonymous Rate (N)	Synonymous Rate (S)	N/S	Non-Synonymous Sites	Synonymous Sites	1st&2nd Codon Sites
ATP8	6	0.4803	0.6786	0.7637	5.45*	0.02	6.95*
ND5	6	0.1682	0.8625	0.1999	2.49	3.41	7.92*
COB	6	0.1342	0.9969	0.1410	19.03**	0.29	19.85***
COII	6	0.1961	1.664	0.1204	46.63**	0.33	31.01***
COI	6	0.0451	1.277	0.0428	12.65**	5.53*	2.94
ND1	5	0.1228	1.572	0.0799	4.40*	1.64	4.23
ND6	4	0.2519	2.165	0.1181			
ND4	4	0.1592	1.844	0.0895			
ATP6	4	0.1312	1.710	0.0854			
ND3	3	0.1822	4.122	0.1941			
COIII	3	0.0621	1.386	0.0523			
ND4L	2	0.1010	0.8720	0.1171			
ND2	0	0.1996	20.31	0.0476			

a - codon model (Muse and Gaut, 1994), nucleotide model (Muse and Weir, 1992). Gene abbreviations are, NADH dehydrogenase subunits 1, 2, 3, 4, 4L, 5 and 6 (ND1, 2, 3, 4, 4L, 5 and 6), ATP synthase subunits 6 and 8 (ATP6 and 8), cytochrome c oxidase subunits I, II, and III (COI, II, and III), and cytochrome b (COB)

* - significant at α=0.05.

** - significant at α=0.01.

*** - significant at α=0.005.

genes also show significant rate differences overall using the nucleotide model. Hence it is unlikely that the differences seen are artefacts introduced by base compositional biases. The COI and ND1 genes also showed a significant non-synonymous rate differences determined with the codon model, but this difference was not present from the nucleotide model. While this may mean that the rate difference seen is due to base compositional bias, it is also likely that any differences in the non-synonymous substitution rate could have been obscured by synonymous changes.

Ratios of synonymous and non-synonymous substitutions

A commonly employed measure used to determine whether adaptive evolution has occurred is the dN/dS ratio. A ratio of greater than one implies strong adaptive evolution has occurred, as opposed to neutral drift. However, sequence comparisons that result in dN/dS ratios less than one do not necessarily mean an absence of selection. In the results presented in Table 4.1, dN/dS ratios were found to much less than one for all mitochondrial genes, except ATP8. Given that the COII and COB genes showed strong evidence of increased non-synonymous substitution rates in simian primates (Adkins and Honeycutt, 1991; Chapter 2), it is interesting that not even these genes showed raised dN/dS ratios. As adaptive evolution usually involves just a small number of non-synonymous changes that have key importance to the function of a protein (see Chapter 1), it would be expected that over larger evolutionary distances small amounts of adaptive non-synonymous change could be obscured by the comparatively huge number of synonymous changes. However, over shorter distances this may not be the case, especially if the lineage where an adaptive change has occurred can be assessed specifically. Through the reconstruction of ancestral sequences it is possible to isolate specific lineages in a phylogeny and estimate branch specific dN/dS ratios. Ancestral sequences are best reconstructed with maximum-likelihood methods (see Yang et al 1995 for the best current technique), but there are statistical problems with using maximum-likelihood estimates of reconstructed ancestral sequences as observed data. The reconstruction method of Yang (1998) is an extension of previous maximum-likelihood methods and allows estimation dN/dS ratios as part of the ancestral

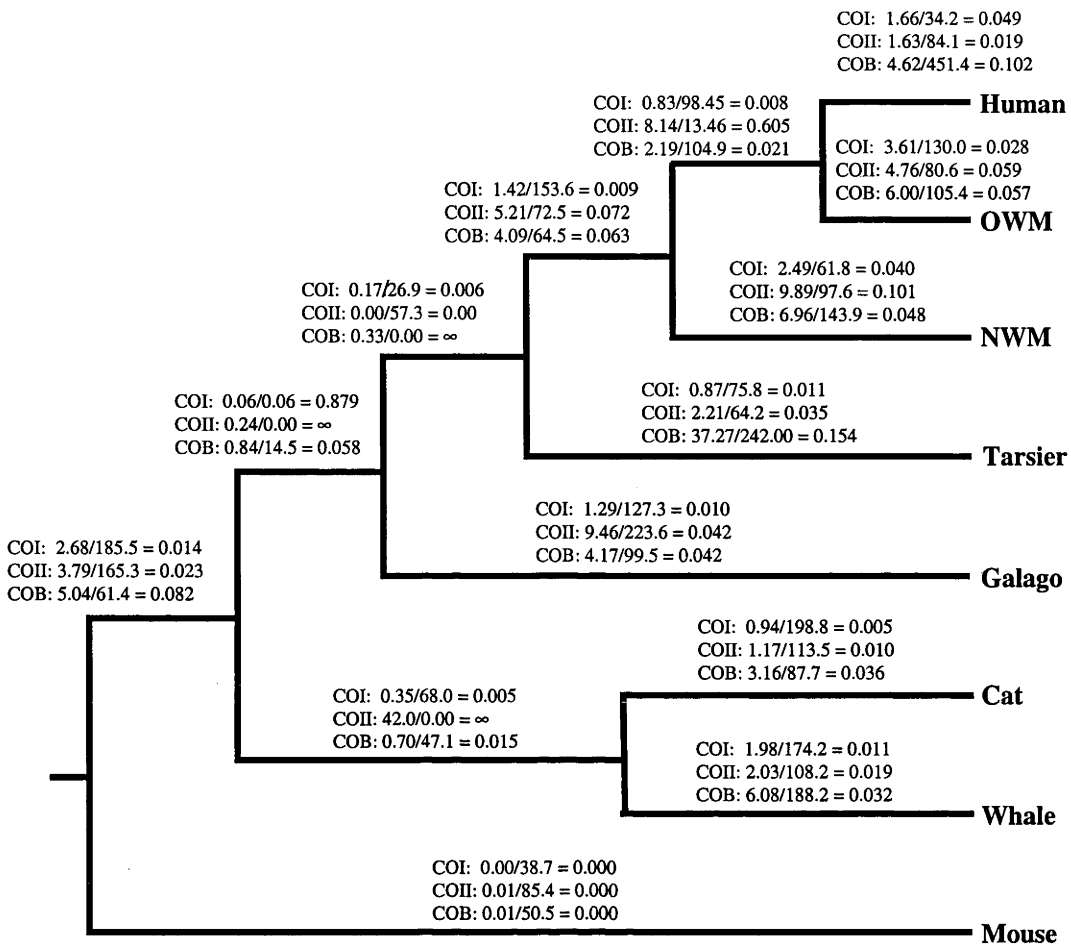


Figure 4.1 - Estimates of non-synonymous and synonymous substitutions per 100 sites, and dN/dS ratios for all branches of phylogenies for the COI, COII and COB genes. OWM and NWM branches refer to branches lead to an Old World Monkey and a New World Monkey species, respectively. The actual species of OWM and NWM were different for each of the three genes (COI, OWM=*Colobus polykomos*, NWM=*Ateles Geoffroyi* ; COII OWM=*Maccaca mulatta* , NWM=*Alouatta paliata* ; COB, OWM=*Colobus guereza*, NWM=*Saimiri scuireus*)

sequence reconstruction process. Here this method has been employed to assess the possibility of elevated rates of non-synonymous substitution (and hence larger dN/dS ratios) in mitochondrial genes along particular lineages leading to apes and other primates.

Figure 4.1 shows a constrained phylogeny of five primates and three other mammals. Along each branch are the maximum-likelihood estimates of the non-synonymous and synonymous substitution rates per hundred sites and the dN/dS ratio for the COI, COII and COB genes. Overall, none of the genes show dN/dS ratios greater than one (excluding two cases of infinite ratios), and in only two instances does the ratio rise above 0.5. A number of lineages shown in Figure 4.1 indicate that more than one hundred substitutions have occurred per hundred synonymous sites, and this compromises estimates of

dN/dS ratios for these lineages. Previous work has shown that each of the three genes have higher non-synonymous substitution rates following the divergence of simians from tarsiers, however on this lineage in figure 4.1 the dN/dS ratios are low and certainly do not imply positive selection. Other branches in this phylogeny do not really show any correlated increases in the dN/dS ratio, although the lineage that separates primates from other mammals shows higher dN/dS ratios for the COI (0.879) and COII (∞) genes, and directly following this the COB gene shows a infinite dN/dS ratio on the lineage leading to haplorhine primates. If it was necessary to ascribe a point where correlated adaptive change had occurred to these genes, potentially a change in the COI and COII genes after the divergence of primates from other mammals was followed by (or triggered) a change in the COB gene on the lineage leading to haplorhine primates. However, such a conclusion is strongly at odds with previous findings.

The hypothesis that each branch of the tree for the COI, COII and COB genes had differing dN/dS ratios was tested against the null hypothesis of uniform dN/dS ratios along all branches. For each gene, the log-likelihood score obtained for the tree created assuming differing dN/dS ratios was lower than that of the tree generated assuming constant dN/dS ratios (data not shown), however, the log ratio test statistics generated from these tests were not significant at a confidence level of α=0.05 with seven degrees of freedom. Testing of various hypotheses of two classes of dN/dS ratios for different branches was not conducted due to the seemingly arbitrary nature of assigning branches to different dN/dS ratio classes.

Discussion

The relative rate tests employed here have shown that a number of genes of the mitochondrial genome display substantial rate heterogeneity between apes and other mammals. The COI, COII and COB genes have been identified previously as having higher non-synonymous evolutionary rates in simian primates than in other mammals (Adkins and Honeycutt 1994; Chapters 2 and 3), and the analysis performed would appear to confirm this. The ATP8, ND1 and ND5 genes were also identified as displaying rate heterogeneities between apes and other mammals, but for these genes the rate differences observed ap-

peared less pronounced than those seen for the COII and COB genes. Three of the genes found here to have higher rates in apes were also found to have rate variations in the analysis of Janke *et al* (1991), where all mitochondrial genes from five mammalian species (importantly including humans) were tested for rate uniformity. It was shown that the COII, COB, COI, ATP6 and ND4 genes (ordered from most significant to least significant, respectively) showed non-uniform evolutionary rates. In the case of the COII gene the rate differences were attributed to the human sequence displaying a higher rate, while the smaller rate variations for the ATP6 and ND4 genes were found only to be present at first or second codon positions. The differences in results found by this study are likely to be due to the differing methodologies used, and that in the present analysis differences in non-synonymous and synonymous rates have been specifically investigated.

It has been shown here that six mitochondrial genes may have undergone an elevation in evolutionary rate, and interestingly these differences have been found to be due to an increase in the rate of non-synonymous substitution. This finding suggests that the rate increases observed for these genes are the result of positive selection. However, the possibility of saturation of synonymous substitutions, which cannot be fully assessed here with existing sequence information, means that it is not possible to say that the increase in non-synonymous substitution rate has not been accompanied by an increase in the synonymous substitution rate. If both the non-synonymous and synonymous substitution rate were raised, then an alternate explanation for this would be an increase in the mutation rate. In Chapter 3, saturation of four-fold degenerate sites in the COI gene was assessed at shorter evolutionary distances than those observed here. While the results of this analysis were not entirely conclusive they alluded that at shorter distances saturation had not occurred, yet raised non-synonymous substitution rates and largely constant synonymous substitution rates were found for simian primate species. Potentially, this may be the case for the other mitochondrial genes shown here to have elevated non-synonymous substitution rates, however, non-simian primate mitochondrial sequences will be required to assess this fully.

Saturation aside, the fact that specific mitochondrial genes have elevated evolutionary rates while others have constant rates implies that the increases observed are not the

result of a genome wide elevation of the mutation rate. This differentiation of rate between genes implies that specific features of the genes with higher non-synonymous substitution rates have caused their rate increase rather than a generic rate increase. Here it is argued that the non-synonymous rate increases are the result of positive selection, however it is not possible to discount other potential causes such as mutational "hot-spots".

If it is assumed that the rate increases seen here for the primate ND1, ND5 and ATP8 sequences are indeed due to positive selection, then what has caused them to be selected? Has this been related to the elevated non-synonymous substitution rates found for the simian COI, COII and COB genes? The increased rates found for the COI, COII and COB genes, which encode components of the cytochrome c oxidoreductase and cytochrome c oxidase complexes (complexes III and IV), have been suggested as being due to co-evolution (coordinated adaptive evolution) of these proteins along with nuclear encoded cytochrome c (Cann *et al* 1984; Chapter 2 and 3). Given that the NADH dehydrogenase complex (complex I; which contains the ND1 and ND5 gene products) and the ATP synthase complex (complex V; which contains the ATP8 gene product) are not closely biochemically associated with either complexes III or IV, and do not bind cytochrome c, it is hard to imagine how the rate accelerations seen for the ND1, ND5 and ATP8 subunits could be related to the postulated coordinated evolution of complex III with complex IV. Most probably the non-synonymous substitution rate increases seen for the ND1, ND5 and ATP8 genes are not related to evolving biochemical interactions between electron transport chain complexes. However, as the ETC is a sequential biochemical pathway, changes in the flux through the pathway due to a change in one complex may have an indirect effect on other complexes. Alternatively, coordinated evolution of these subunits with a nuclear encoded protein with the NADH dehydrogenase complex could be the reason for these rate increases. Access to a larger number of primate sequences for these genes will provide more information on this matter, especially as this will make it possible to determine if the species range for which the non-synonymous rate increases are seen is the same as that of the COI, COII and COB genes.

Perhaps the most curious finding of this work has been that the estimation of synonymous and non-synonymous ratios for each branch of a phylogeny using the method of

--- ADDENDUM ---

(Ch 4 and 5 - general comment on methodology) - This thesis concentrated on a number of relatively large rate differences that exist between certain mitochondrial genes of simian primates and other mammals. The tests presented in this thesis largely show there to be little evolutionary rate difference between non-primate mammal sequences, however some differences did exist. In all cases, rate differences between the non-primate mammal sequences were substantially smaller in magnitude (as evidenced by log likelihood ratios and comparison of K values). In one particular case in chapter 4, a non-synonymous rate increase was evident for a number of genes from the whale mitochondrial genome (when these were compared horse and cat genes - data not shown, although see data appendix). The rate increase was not present for the same genes as were shown to have an increased rate in simian primates, and it was hard to imagine that this rate fluctuation could be connected to a potential adaptive evolutionary episode in the whale ETC (because the effect was so small). However, this result raises the question of whether this was another adaptive phenomena, or just an artefact of the data? Are small fluctuations in evolutionary rate commonplace, and how big does a rate fluctuation have to be to be considered adaptive? Answering this question is difficult and is one of the current challenges of this kind of study. More specifically, how can adaptive evolution be appraised when K_a/K_s ratios between candidate sequences are less than one, especially when it is noted that adaptive evolution can occur through a very small number of changes which will not raise a K_a value to being greater than K_s?

Yang (1998) did not show the increased rates of non-synonymous substitution found to be evident for a number of the mitochondrial genes (namely, COI, COII and COB). In the cases of the COII and COB genes, the very large elevation on non-synonymous substitution rate in simian primates shown by previous studies (Adkins and Honeycutt, 1992; Chapter 2) were not detectable. Increases in dN/dS ratio were evident for all three genes on certain lineages, but these did not correlate with what had been previously found to be the case. Potentially, if synonymous substitutions were saturated on various lineages this would affect the dN/dS ratio calculated, but it would be expected that this would actually raise the ratio. However, saturation of synonymous sites may interfere with accuracy of the likelihood maximisation process in a counter-intuitive manner. Rather than looking for high dN/dS ratios on specific branches, a more useful analysis could be to compare the ratios on each lineage against a number of other mitochondrial genes that have been shown here to have uniform evolutionary rates between mammals and primates. With this information, a series of non-parametric sign tests could be employed in a similar manner to those performed above for the maximum-likelihood relative rate tests. As more primate mitochondrial sequence information becomes available this will be an interesting analysis to explore.

Chapter Five

Estimation of divergence dates among apes and other primates using a mitochondrial gene dataset free of rate and base compositional heterogeneities.

Abstract

Previous studies of divergence dates among apes and other primates using mitochondrial genes have been conducted without proper regard for evolutionary rate and base compositional differences known to exist between mammals. Here a subset of genes of the mitochondrial genome which show a lack of rate heterogeneity between primates and other mammals have been used to estimate divergence dates. Another dataset of ND2 genes from a larger range of primates has also been used to estimate divergence dates among the primates. The distance matrices estimated from each of these datasets imply a smaller range of divergence dates among mammals than predicted by the fossil record. Distance estimates between species were scaled to evolutionary time both with a range of fossil dates as well as by applying different substitution rates. While acceptable sets of divergence dates could be resolved separately for apes, for non-ape primates and for deeper mammalian divergences, no set of dates could be derived that was satisfactory for all species.

Introduction

Using molecular information to resolve the divergence dates of mammalian species has been a pastime enthusiastically pursued throughout the history of the field of molecular evolution. Traditionally this has been a highly controversial area of study, reflecting the difficulties that are encountered in gaining reliable molecular distances and adapting these to evolutionary time, and how matters of interpretation can greatly influence results obtained. However, with the explosion of sequence information becoming available in the genome databases and the development of robust statistical models of sequence evolution, larger datasets have been used to obtain reliable estimates of genetic distance between species, and a number of matters of long-standing controversy have been largely solved. Of particular note, the divergence order among the hominoids has been hotly contested for over a decade, however consensus has recently been reached and hominoid inter-relationships are now believed to be ((Homo,Pan),Gorilla) (Ruvolo 1997). An actual consensus on the divergence dates between these species has not been fully agreed upon, and the reader is referred to a number of studies (which have analysed either mitochondrial or nuclear sequences, or both) which represent a range of current thinking (Hasegawa et al 1985; Hasegawa et al 1989; Kishino and Hasegawa 1990; Hasegawa 1990; Horai *et al* 1992; Adachi and Hasegawa 1995; Easteal *et al* 1995, p126; Horai *et al* 1995; Easteal and Herbert 1997; Kumar and Hedges 1998).

More recently, the steady increase in the number of wholly sequenced mitochondrial genomes has allowed consideration of divergences within the apes through comparison of sequences of all mitochondrial protein-coding genes (Arnason *et al* 1996). However, instead of supporting or refining the dates obtained by previous studies, the dates presented by this study are much higher than what is considered possible given the fossil record, and once again debate has been re-opened on this subject. Arnason *et al* (1996) used a divergence between artiodactyls and cetaceans dated at around 60 million years (Arnason *et al* 1996; Kumar and Hedges 1998) to calibrate the inter-species evolutionary distances estimated from the concatenated sequences of all mitochondrial protein-coding genes. By using the divergence of artiodactyls and cetaceans to calibrate primate divergences (instead of using a primate divergence date to calibrate other primate divergences),

--- ADDENDUM ---

(p75 , para 1, replace first sentence with) -- Arnason *et al* have calculated dates that are derived from independent lineages, and these could be expected to be better than those obtained previously (see Hillis 1996).

see addendum

~~Arnason *et al* have obtained dates that for independent lineages, and could be expected to be better than those previously obtained (see Hillis *et al* 1996).~~ While use of independent lineages to calibrate genetic distance to time is a better approach to determining divergence dates, the greater distances involved lay the results of such an analysis prone to inaccuracies. The divergence dates of Arnason *et al* have been strongly criticised for failing to adequately account for evolutionary rate and base compositional heterogeneities known to exist between the mitochondrial genes of primates and other mammals (Penny *et al* 1998).

Evolutionary rate heterogeneity between primates and other mammals has been shown to exist for a number of mitochondrial genes, proposed to be due to an episode of adaptive co-evolution between some genes in simian primates (Adkins and Honeycutt 1994; Chapters 2, 3 and 4). However, lack of rate heterogeneity between mammals has also been observed for other mitochondrial genes (Chapter 4). Base compositional differences have also been shown to be strong between mammalian species, but this has been shown for a number of genes to be largely due to heterogeneity present at third codon positions (Chapters 2 and 3). Here, using a dataset of mitochondrial genes which do not display rate variation between primates, artiodactyls and cetaceans, and for which base compositional differences have been addressed, divergence dates between the apes and other primates have been re-estimated.

Materials and Methods

Sequencing and Data Sources. Liver samples from *Tarsius syrichta* (Philippine Tarsier) and *Galago senegalensis* (lesser bushbaby) were obtained from the Duke University Primate Centre, a blood serum sample from *Ateles geoffroyi* (spider monkey) was obtained from the Royal Melbourne Zoological Gardens and a blood-clot sample from *Cercopithecus aethiops* (African green monkey) was were supplied by D. C. Gajdusek and C. J. Gibbs. All DNA samples were obtained by phenol/chloroform extraction (Sambrook *et al* 1989) from the original tissue, which in the case of the liver samples had been homogenised by grinding with a mortar and pestle after being frozen with liquid nitrogen. The ND2 genes were isolated using the polymerase chain reaction with a pair of "universal"

primers, as shown in Table 5.1. The single product resulting from each polymerase chain reaction was purified using Wizard PCR preps (Promega), and subsequently sequenced using a dye-terminator cycle-sequencing protocol (Perkin-Elmer) with the universal primers and a series of internal sequencing primers (Table 5.1) on an ABI377 automated sequencer (Applied Biosystems). Sequences will be submitted to the DDBJ/EMBL/GenBank databases in the near future. Sequences of complete mitochondrial genomes were obtained from published sources, *Homo sapiens* (human, J01415) (Anderson *et al* 1981), *Pan troglodytes* (chimpanzee, D38113), *Gorilla gorilla* (gorilla, D38114), *Pongo pygmaeus* (orangutan, D38115) (Horai *et al* 1995), *Hylobates lar* (gibbon, X99256) (Arnason *et al* 1996), *Balaenoptera physalus* (whale, X61145) (Arnason *et al* 1991), *Bos taurus* (cow,) (Anderson *et al* 1982, V00654), *Mus musculus*, (mouse, J01420) (Bibb *et al* 1981), *Didelphis virginiana* (opossum, Z25973) (Janke *et al* 1994), and the sequences of individual protein-coding genes edited from these.

Data Analysis. Two alignments of mitochondrial sequences were created, one consisting of only ND2 sequences and another of a set of concatenated mitochondrial genes. The inferred amino acid sequences of both sets of genes were aligned using CLUSTALW (Thompson *et al* 1994), and any gaps inserted in this alignment were transferred to an alignment of nucleotide sequences. In this manner, frameshifts in the aligned sequences were avoided. Potential base compositional heterogeneity between the sequences of each alignment was assessed using the Distance program by Jermiin *et al* (1998). Using this method, pairwise Z-statistics were calculated at each codon position between all sequences, and the results plotted graphically to allow comparison of values between codon positions. Uniformity of evolutionary rates between sequences was assessed using the maximum-likelihood relative ratio test of Muse and Weir (1992). Evolutionary distance estimated by the methods Tamura and Nei (1993), Jukes and Cantor (1969) and Kimura (1980) was calculated using the MEGA software package (Kumar *et al* 1993). Evolutionary distance estimated by the method of Hasegawa *et al* (1985) was calculated using the Puzzle program (version 4.0) by Strimmer and von Haeseler (1996). Non-degenerate distances were calculated with the method of Wu and Li (1985).

Table 5.1 - Oligonucleotide primers used for the amplification and sequencing of ND2 genes.

Species	Primer[b]	Sequence (5'-3')[c]
Universal primers*a*	H4373	GGCCCATACCCCGAAAATGTT
	L5558	TICTIAGGGCTTTGAAGGC
Cercopithecus aethiops	H4691	TCCTCCTCACATGAMARAAA
	H4735	ACCACCAATCAACTCCCAT
	H5131	CTAAAITCAAACACCACAAC
	L4882	GGGCTAGTTTYTGTCATGT
	L5252	ACAAAGCCGGTCAGTGGA
Ateles geoffroyi	H4861	CTAACATGACAIAAACTAGC
	H5098	TCCTAACAATCTCTACACTC
	H5223	AGGAGGTITACCYCCRCT
	L4882	as above
	L5028	TATAGGTGATTGAGGAGTAA
	L5256	TAGGGGGAAAAGCCTGTTA
Tarsius syrichta	H4691	as above
	H5216	TATCMCTAGGAGGACTICC
	L4882	as above
	L5234	GGTAGTCCTCCTAGIGATA
Galago senegalensis	H4859	CCTAACATGACAAAAAYTAGC
	H5216	as above
	L5234	as above
	L4883	CGGGCTARTTTTTGTCAT
	L5147	GTGGTGGTATTAGAGTTTGA

a - universal primers were used for the first sequencing steps from both ends of all genes.

b - primer numbers refer to the nucleotide at the 5' most end of the sequence and are numbered with reference to the scheme of Anderson *et al* (1981). H and L refer to the strand to which the primer anneals.

c - I represents deoxyinosine, degenerate bases are represented by standard notation.

Results

Many mitochondrial genes have been found to demonstrate large evolutionary rate differences between simian primates and other mammalian species (Adkins and Honeycutt 1994; Chapters 2, 3 and 4). Assigning dates to divergences between sequences that have evolved at different rates since their divergences is difficult, and a theoretical framework for doing this has not yet been fully established, although some new methods have recently appeared (Adachi and Hasegawa 1995; Rambaut and Bromham 1998). Hence, estimation of divergence dates among species from nucleotide data can be greatly simplified by choosing a dataset that contains relatively little rate heterogeneity between sequences. Previous work has identified primate mitochondrial genes which show rate differences between primates and other mammals (Chapter 4), and here they are excluded from this dataset of protein-coding genes. The primary dataset for the following analysis of evolutionary distances between primate species is that of the NADH dehydrogenase subunit 2 gene. This gene has demonstrated a lack of rate heterogeneity between mammalian species, and is ideal for a comparative study of this nature. As well as this dataset, an additional concatenation of the NADH dehydrogenase subunits 2, 3, 4, 4L, 6, ATP synthase subunit 6 and cytochrome c oxidase subunit III genes (ND2, ND3, ND4, ND4L, ND6, ATP6 and COIII) has been used. The second dataset consists of a smaller range of species and comprises all mitochondrial genes which did not show rate heterogeneity in the previous study (Chapter 4). The dataset of concatenated genes overcomes a deficiency of the ND2 dataset, in that the sequences analysed are of much greater length.

Base compositional differences between species for both datasets were assessed in a pairwise manner, and a *Z*-statistic representing the degree of base compositional difference between each pair of sequences calculated. As shown in figure 5.1, each codon position was compared separately, and for both datasets the magnitude and range of *Z*-statistics is greater for third codon positions than for both first and second codon positions. This demonstrates that there is relatively little base compositional heterogeneity between sequences at the first and second codon positions, but substantial heterogeneity between third codon positions. In order to obtain a dataset mostly free of base compositional differences, third codon positions have been excluded from the subsequent analysis.

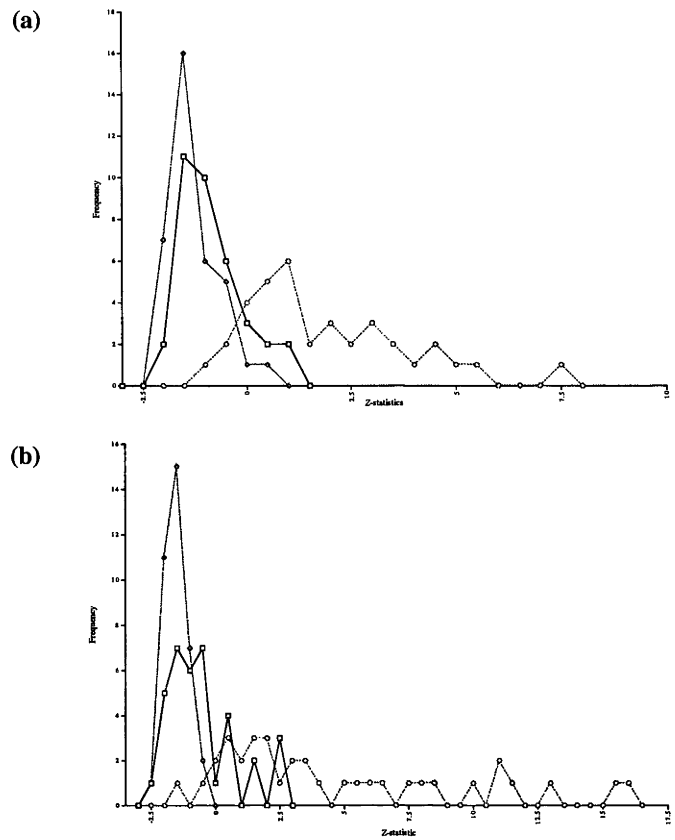


Figure 5.1 - Frequency distribution of Z-stattistics for pairwise comparisons between (a) ND2 sequences and (b) the sequences of the concatenated genes. Squares show first codon sites; diamonds show second codon sites; and circles, third codon sites.

To check that rate heterogeneity was absent from the dataset of first and second codon positions, relative rate tests were performed on each dataset and likelihood ratios calculated for comparisons between the primate, whale and cow sequences, using the mouse sequence as an outgroup. Additional tests were conducted between the primate sequences using more closely related species as outgroups. The results of these tests are shown in Table 5.2. Most comparisons do not show significant rate heterogeneity (tested with log-likelihood ratios which approximate a χ^2 distribution with two degrees of freedom) between primates and other mammals, and hence estimates of evolutionary distance made between the cow and the whale sequences (the artiodactyl-cetacean divergence) should be directly comparable to distances estimated between the primate species. However some rate differences do exist between the apes for both the ND2 and concatenated datasets. Given that a large number of multiple comparisons performed here, these sig-

nificant log ratios could be type-I errors. After Bonferroni correction (Rice 1989) of Table 5.2, only the differences between the human and gorilla, and the human and orangutan concatenated sequences dataset were found to be significant. While the Bonferroni correction yields conservative results when the data being tested are not independent, such as in a relative rate test, the results imply that the ape divergence dates may be best estimated using the ND2 dataset.

Table 5.2 - Rate differences between species appraised using the maximum-likelihood method of Muse and Weir (1992)[a].

Outgroup	Species 1	Species 2	ND2	Concatenated Genes
Mouse	Whale	Cow	0.46	4.42
		Human	0.14	4.96
		Chimpanzee	0.07	6.26*
		Gorilla	0.49	4.51
		Orangutan	0.81	2.84
		Gibbon	0.01	1.87
		African Green Monkey	1.45	-
		Spider Monkey	1.64	-
		Tarsier	0.02	-
		Galago	2.25	-
Mouse	Galago	Human	1.57	-
		Chimpanzee	1.70	-
		Gorilla	2.65	-
		Orangutan	3.26	-
		Gibbon	0.84	-
		African Green Monkey	2.12	-
		Spider Monkey	2.72	-
		Tarsier	3.78	-
Galago	Tarsier	Human	4.68	-
		Chimpanzee	4.16	-

(continued)

Table 5.2 - continued.

Outgroup	Species 1	Species 2	ND2	Concatenated Genes
		Gorilla	3.21	-
		Orangutan	1.80	-
		Gibbon	5.74	-
		African Green Monkey	4.23	-
		Spider Monkey	2.61	-
Tarsier	Spider Monkey	Human	3.95	-
		Chimpanzee	4.19	-
		Gorilla	2.55	-
		Orangutan	2.37	-
		Gibbon	4.19	-
		African Green Monkey	2.91	-
Spider Monkey	African Green Monkey	Human	4.19	-
		Chimpanzee	5.32	-
		Gorilla	3.39	-
		Orangutan	1.30	-
		Gibbon	4.69	-
African Green Monkey	Gibbon	Human	7.00*	-
		Chimpanzee	5.57	-
		Gorilla	5.85	-
		Orangutan	0.55	-
Whale	Gibbon	Human	-	2.20
		Chimpanzee	-	0.13
		Gorilla	-	1.06
		Orangutan	-	0.45
Gibbon	Orangutan	Human	7.63*	7.62*
		Chimpanzee	9.13*	9.36**
		Gorilla	5.90*	21.72***
	Gorilla	Human	0.87	18.68***
		Chimpanzee	0.59	7.37*
	Chimpanzee	Human	1.84	2.20

a - results of tests represented by log-likelihood ratios, $2(l_i-l_0)$. Significant difference in rate between species at a confidence level of (*) α=0.05 (**) α=0.01 (***) α=0.005 with 2 d.f.

The genetic distance estimates made using the method of Tamura and Nei (1993) are shown in Tables 5.3a and 5.3b. Distances were also estimated using the method of Jukes and Cantor (1969), Kimura's two-parameter model (1980), and the method of Hasegawa *et al* (1985) and the matrices obtained for each dataset for each model were very similar (data not shown).

The distance estimates shown in Tables 5.3a and 5.3b were used with the fossil derived 60Ma divergence date of artiodactyls and cetaceans to estimate the divergence dates among apes and other primates (Table 5.4a). The ape divergence dates obtained using this calibration point for both the concatenated and ND2 datasets are even greater then those arrived at by Arnason *et al* (1996). Table 5.4a also shows divergence dates that result when inter-specific genetic distances are calibrated with other commonly used fossil dates. The dates obtained using a 14Ma divergence date between the African apes and the orangutan give more recent primate divergences, but older divergences between mammalian orders also become substantially less ancient. A calibration date of 65Ma for the divergence of primates and artiodactyls/cetaceans results in a median set of dates where the ape divergence dates are older and the deeper divergences among the mammals are younger.

As an alternative to using fossil calibration dates to convert genetic distance to time, which leads to divergence time estimates that suffer from a lack of independence when primate dates are used as calibration points, divergence dates can be calculated by applying a range of different substitution rates to estimates of distance (see Easteal and Herbert (1997) and Easteal, Collet and Betty (1995), p126). Table 5.4b shows the range divergence dates calculated for a range of assumed rate constants for both the ND2 and concatenated gene datasets. Divergence dates were calculated for substitution rates ranging from 1.0 to 6.0 $x10^{-9}$ substitutions per site per year, this range encompassing the previous estimate of Horai *et al* (1995) for the substitution rate of mammalian mitochondrial DNA ($3.5x10^{-9}$ substitutions per site per year).

The range of dates calculated here for mammalian divergences may be affected by the saturation of synonymous codon sites, as suspected from previous chapters. Hence,

Table 5.3(a) - Tamura and Nei (1993) distances between ND2 genes[a].

	Hsa	Ptr	Ggo	Ppy	Hla	Cae	Age	Tsy	Gse	Bta	Bph	Mmu	Dvi
Hsa	0	0.85	1.09	1.54	1.48	2.06	2.32	2.53	2.51	2.32	2.43	2.76	3.03
Ptr	4.68	0	0.83	1.38	1.31	2.21	2.33	2.62	2.55	2.40	2.49	2.83	3.08
Ggo	7.25	4.27	0	1.39	1.34	2.19	2.40	2.67	2.58	2.54	2.49	2.90	3.04
Ppy	12.52	10.89	10.90	0	1.68	2.37	2.53	2.70	2.50	2.59	2.65	2.92	3.20
Hla	13.27	10.11	10.47	15.05	0	2.45	2.33	2.75	2.71	2.59	2.65	2.80	3.21
Cae	20.46	23.03	22.58	25.23	26.52	0	2.52	3.11	2.63	2.62	2.56	2.93	3.05
Age	25.29	26.13	27.17	29.07	26.44	28.66	0	2.77	2.43	2.53	2.64	2.84	3.49
Tsy	30.13	32.04	32.89	33.43	34.20	37.81	34.55	0	2.27	2.36	2.53	2.72	2.79
Gse	30.45	31.32	31.56	30.21	33.56	31.89	29.58	26.32	0	2.20	2.41	2.67	2.86
Bta	27.23	28.74	30.62	32.48	31.90	31.81	31.27	28.55	26.03	0	1.58	2.70	2.87
Bph	28.57	29.84	29.60	31.61	32.74	30.03	32.83	30.69	28.76	14.46	0	2.73	2.87
Mmu	35.06	36.64	37.59	37.88	36.51	37.83	37.38	34.89	34.44	34.48	34.83	0	3.04
Dvi	40.02	41.04	40.20	43.08	43.38	39.23	47.59	37.15	38.32	38.04	37.73	41.61	0

a - below diagonal are estimates of distance x100, above diagonal and the standard error x 100 associated with these estimates.

Table 5.3(b) - Tamura and Nei (1993) distances between concatenated genes[a].

	Hsa	Ptr	Ggo	Ppy	Hla	Bta	Bph	Mmu	Dva
Hsa	0	0.41	0.42	0.62	0.58	0.87	0.91	0.95	1.03
Ptr	5.40	0	0.43	0.61	0.57	0.88	0.89	0.97	1.03
Ggo	5.43	5.59	0	0.58	0.56	0.87	0.87	0.96	1.04
Ppy	11.01	10.64	9.73	0	0.64	0.88	0.87	0.94	1.08
Hla	9.75	9.33	8.94	11.48	0	0.91	0.88	0.94	1.02
Bta	20.72	21.12	20.75	20.82	21.79	0	0.67	0.88	1.01
Bph	21.75	21.23	20.37	20.49	20.88	13.20	0	0.92	1.06
Mmu	24.17	24.80	24.37	23.88	23.81	21.47	22.89	0	1.01
Dvi	27.71	27.55	27.87	29.45	27.23	27.02	28.75	27.20	0

a - below diagonal are estimates of distance x 100, above diagonal and the standard error x 100 associated with these estimates.

the above estimation of divergence dates was repeated using distances calculated at non-degenerate sites. However, the results obtained were not substantially different to those presented in Tables 5.3 and 5.4 (data not shown).

Discussion

This study has assessed divergence dates between apes and other primates using a dataset of mitochondrial genes shown to be relatively free of both rate and base compositional heterogeneity. This data is a subset of that used by Arnason *et al* (1996), and using the same calibrating fossil date of 60Ma for the divergence of Artiodactyls and Cetaceans, even larger divergence dates than those found previously were calculated among the apes. Arnason *et al* did not rigorously consider potential evolutionary and base compositional heterogeneities between species in their analysis, and the unusually old divergence dates they obtained for apes have been dismissed as resulting from a higher evolutionary rate in primates compared than in other mammals (Penny *et al* 1998). However, here it is shown that once rate and base compositional heterogeneity is removed from the nucleotide data, the divergence dates among primates are still extremely old when the 60Ma artiodactyl/ cetacean calibration date is used. When other commonly used fossil calibration dates are applied, divergence dates among the primates become locally acceptable but other more

Table 5.4(a) - Relative evolutionary distances, and inferred divergence dates.

Divergence[a]	ND2 Genes - Divergence (Ma)				Concatenated Genes - Divergence (Ma)			
	Distance±SE	1	2	3	Distance±SE	1	2	3
Homo - Pan	4.68±0.85	19.4	5.2	10.8	5.40±0.41	24.5	6.9	16.7
Homo - Gorilla	7.25±1.09	30.1	8.1	16.8	5.43±0.42	24.7	6.9	16.8
Homo - Pongo	12.52±1.54	52.0	[14][b]	29.0	11.01±0.62	50.0	[14][b]	34.1
Homo - Hylobates	13.27±1.48	55.1	14.8	30.8	9.75±0.58	44.3	12.4	30.2
Hominoidea - Cercopithecoidea	20.46±2.06	84.9	22.7	47.4	-	-	-	-
Catarrhini - Platyrrhini	25.29±2.32	104.9	28.3	58.6	-	-	-	-
Simiformes - *Tarsius*	30.13±2.53	125.0	33.7	69.8	-	-	-	-
Haplorhini - Strepsirhini	30.45±2.51	126.3	34.0	70.6	-	-	-	-
Artiodactyla - Cetacia	14.46±1.58	[60][b]	16.2	33.5	13.20±0.67	[60][b]	16.8	40.9
Primates - Artiodactyla/Cetacia	28.05±2.38	116.4	31.4	[65][b]	20.99±0.88	95.4	26.7	[65][b]
Primates - Rodentia	35.06±2.76	145.5	39.2	81.2	24.21±0.95	110.0	30.8	75.0
Rodentia - Marsupials	41.61±3.04	172.7	46.5	96.4	27.20±1.01	123.6	34.6	84.2

a - the representative species for Homonoidea, Simiformes, Haplorhini and Primates is *Homo sapiens*, for Cercopithecoidea and Catarrhini is *Cercopithecus aethiops*, for Platyrrhini is *Ateles geoffroyi*, for Strepsirhini is *Galago senegalensis*, for Artiodactyla is *Bos taurus* , for Cetacia is *Balaenoptera physalus*. for Rodentia is *Mus musculus* and Marsupials is *Didelphis virginiana*.

b - calibration dates taken from fossil evidence.

Table 5.4(b) - Divergence dates given different assumed substitution rates.

Divergence[a]	ND2 Genes - Divergence (Ma)						Concatenated Genes - Divergence (Ma)					
Substitution rate (x10^9)	1.0[b]	2.0	3.0	4.0	5.0	6.0	1.0[b]	2.0	3.0	4.0	5.0	6.0
Homo - Pan	23.4±4.3	11.7±2.1	7.8±1.4	5.9±1.1	4.7±0.9	3.9±0.7	27.0±2.1	13.5±1.0	9.0±0.7	6.8±0.5	5.4±0.4	4.5±0.3
Homo - Gorilla	36.3±5.5	18.1±2.7	12.1±1.8	9.1±1.4	7.3±1.1	6.0±0.9	27.2±2.1	13.6±1.1	9.1±0.7	6.8±0.5	5.4±0.4	4.5±0.4
Homo - Pongo	62.6±7.7	31.3±3.9	20.9±2.6	15.7±1.9	12.5±1.5	10.4±1.3	55.1±3.1	27.6±1.6	18.3±1.0	13.8±0.8	11.0±0.6	9.2±0.5
Homo - Hylobates	66.4±7.4	44.2±3.7	22.0±2.5	16.6±1.9	13.3±1.5	11.1±1.2	48.8±2.9	24.4±1.5	16.3±1.0	12.2±0.7	9.8±0.6	8.1±0.5
Hominoidea - Cercopithecoidea	102.3±10.3	51.2±5.2	34.1±3.4	25.6±2.6	20.5±2.1	17.1±1.7	-	-	-	-	-	-
Catarrhini - Platyrrhini	126.5±11.6	63.2±5.8	42.2±3.9	31.6±2.9	25.4±2.3	21.2±1.9	-	-	-	-	-	-
Simiformes - *Tarsius*	150.7±12.7	75.3±6.3	50.2±4.2	37.7±3.2	30.1±2.5	25.1±2.1	-	-	-	-	-	-
Haplorhini - Strepsirhini	152.3±12.6	76.1±6.3	50.8±4.2	38.1±3.1	30.5±2.5	25.4±2.1	-	-	-	-	-	-
Artiodactyla - Cetacia	72.3±7.9	36.2±4.0	24.1±2.6	18.1±2.0	14.5±1.6	12.1±1.3	66.0±3.4	33.0±1.7	22.0±1.1	16.5±0.8	13.2±0.7	11.0±0.6
Primates - Artiodactyla/Cetacia	140.3±11.9	70.1±6.0	46.8±4.0	35.1±3.0	28.1±2.4	23.4±2.0	105.0±4.4	52.5±2.2	35.0±1.5	26.2±1.1	21.0±0.9	17.5±0.7
Primates - Rodentia	175.3±13.8	87.7±6.9	58.4±4.6	43.8±3.5	35.1±2.8	29.2±2.3	121.0±4.8	60.5±2.4	40.3±1.6	30.3±1.2	24.2±1.0	20.2±0.8
Rodentia - Marsupials	208.1±15.2	104.0±7.6	69.4±5.1	52.0±3.8	41.6±3.0	34.7±2.5	136.0±5.1	68.0±2.5	45.3±1.7	34.0±1.3	27.2±1.0	22.7±0.8

a - the representative species for Homonoidea, Simiformes, Haplorhini and Primates is *Homo sapiens*, for Cercopithecoidea and Catarrhini is *Cercopithecus aethiops*, for Platyrrhini is *Ateles geoffroyi*, for Strepsirhini is *Galago senegalensis*, for Artiodactyla is *Bos taurus* , for Cetacia is *Balaenoptera physalus*. for Rodentia is *Mus musculus* and Marsupials is *Didelphis virginiana*.

b - divergence dates calculated by (0.5 distance/substitution rate).

distant divergences become too recent. Furthermore, when nucleotide substitution rates are used to calibrate genetic distances rather than fossil dates, no single rate can be shown to be appropriate for all groupings of species. While a higher substitution rate of around $4\text{-}5\text{x}10^{-9}$ substitutions per site per year resulted in a series of likely divergence dates for apes, the dates for other primates and other mammals became far too recent. Likewise, when a lower substitution rate of around $1\text{-}2\text{x}10^{-9}$ substitutions per site per year was applied the dates for deep mammalian divergences became reasonable, but subsequently the divergence dates among primates (especially between the apes) were much too ancient.

From these results it can be seen that the genetic distances calculated between species do not reflect what is believed to be the case from fossil dates. Clearly, the range of dates derived from the fossil record is much larger than that implied by the distance matrix calculated from these mitochondrial genes. Hence, if it is unlikely that any single set of dates presented in Tables 5.4a or 5.4b are actually the true divergence dates between the primate and other mammalian species, what has been the cause of this inability to derive realistic divergence times? While it is possible that the fossil data has been misinterpreted, previously results such as those presented here have been explained as being due to mitigating factors present in the nucleotide data, such as rate and base compositional heterogeneities. However, this explanation can not be applied here as the data used was mostly free of these differences between species. While it is possible that other unidentified confounding features of the nucleotide dataset have not been recognised and compensated for, there may be other explanations for the result of this analysis. The cause of the unlikely dates calculated here is most probably a combination of all or some of the following.

Most prominently, saturation of synonymous substitutions would cause a narrower range of divergence estimates than what was actually the case, and previous chapters have not been able to rule out the possibility that this may have occurred. A set of dates were re-estimated using a dataset of non-degenerate substitutions, which from previous work did no appear to be saturated, however, the dates obtained from this were did not display a greater range than that found using all sites. Hence, while saturation of synonymous

substitution may have caused some reduction of the range of divergence estimates calculated here, this has not been a major factor in producing the unusual dates found here.

This study did not statistically appraise differences of evolutionary rates between the opossum and the mouse, and this may have allowed the introduction of small systematic errors. While rate differences between the opossum and mouse species would not have effected estimates of evolutionary distance between the primates or the other mammals, they could have affected the primate/rodent, rodent/marsupial divergence estimates. Potentially this may partially explain why it was not possible to find a universally acceptable substitution rate for all species in Table 5.4b. However, the distance matrix presented in Tables 5.3a and 5.3b do not allude to any great evolutionary rate differences between the primate, rodent and marsupial species, and any inaccuracy introduced by this could only be small.

A factor that could have produced a misleadingly narrow range of divergences, but would not have been detected by the analysis conducted here, would have been an overall increase in the evolutionary rate of all mammals. Relative rate tests were used to identify rate differences between species, but they do not imply constancy of rate over long periods of evolution. Use of multiple fossil dates with measures of genetic divergence (temporal scaling) has been proposed as a way to investigate rate constancy over large periods of evolution (Gingerich 1986). While this is outside the scope of this chapter, full use of fossil information with a larger dataset of mammalian sequences may allow appraisal of an overall increase in mammalian mitochondrial evolutionary rates, and whether this has been the cause of the narrow divergence estimates found here and by Arnason *et al* (1996).

Perhaps the biggest problem encountered in this study was that after rate and base compositional heterogeneities were removed from the mitochondrial genome, the remaining dataset on which the analysis was conducted was quite small. Certainly, the contradictions present in the results obtained using the ND2 gene data can probably be attributed to the fact that this dataset was for a single gene of only 686bp. Hence it is possible that the evolutionary story told by this single gene may be biased and may not reflect the true evolutionary history of the species analysed here. While the concatenated gene data was more substantial, after third codon positions were removed it still only comprised 3344bp

--- ADDENDUM ---

(Ch 4 and 5 - general comment on methodology) - This thesis concentrated on a number of relatively large rate differences that exist between certain mitochondrial genes of simian primates and other mammals. The tests presented in this thesis largely show there to be little evolutionary rate difference between non-primate mammal sequences, however some differences did exist. In all cases, rate differences between the non-primate mammal sequences were substantially smaller in magnitude (as evidenced by log likelihood ratios and comparison of K values). In one particular case in chapter 4, a non-synonymous rate increase was evident for a number of genes from the whale mitochondrial genome (when these were compared horse and cat genes - data not shown, although see data appendix). The rate increase was not present for the same genes as were shown to have an increased rate in simian primates, and it was hard to imagine that this rate fluctuation could be connected to a potential adaptive evolutionary episode in the whale ETC (because the effect was so small). However, this result raises the question of whether this was another adaptive phenomena, or just an artefact of the data? Are small fluctuations in evolutionary rate commonplace, and how big does a rate fluctuation have to be to be considered adaptive? Answering this question is difficult and is one of the current challenges of this kind of study. More specifically, how can adaptive evolution be appraised when K_a/K_s ratios between candidate sequences are less than one, especially when it is noted that adaptive evolution can occur through a very small number of changes which will not raise a K_a value to being greater than K_s?

for each species. Although for the concatenated genes this was an acceptably sized dataset, it may not have been sufficient for it to have been unaffected by major gene-specific biases. Once again, the evolutionary distances estimated from this data may not have been representative of the actual evolutionary distances.

Overall, the tight range of evolutionary distances found between the primate and other mammalian species was probably due to a combination of the above factors. From here, a greater understanding of recent divergences between primates from mitochondrial sequences might be gained through the use of methods which rigorously compensate for rate heterogeneities between sequences. In employing these methods, larger sequence datasets can be used and potentially more reliable estimates of genetic distance can be made. Also, as the number of complete mammalian mitochondrial genome sequences present in the gene databases increases, denser mammalian datasets can be used to gain better understanding of phylogenetic relationships and evolutionary rate differences between mammals, and more robust calibrations of genetic distance will be achieved. Parallel studies of this kind using both mitochondrial and nuclear sequences perhaps show the greatest promise in accurately estimating primate divergence dates, once these combined datasets become sufficiently large and consist of a diverse range of genes.

--- ADDENDUM ---

(p 90, para 2, after second sentence) -- (although please note that even though it is assumed otherwise in the following discussion, the possibility of a relaxation of functional constraint rather than adaptive selection could not be ruled out - see relevant sections in chapters 2 and 3).

Chapter Six

General Discussion and Future Directions

The major finding of this work has been the identification of a potential episode of adaptive evolution of the electron transport chain (ETC) of simian primates. This adaptive episode appears to have involved a number of enzymatic components of at least two complexes of the ETC, and has involved both nuclear and mitochondrial genes. The mitochondrial gene encoding cytochrome b (COB), a major electron carrying subunit of complex III, has been found to have an accelerated evolutionary rate in simian primates and this increase can be attributed largely to an increase in the rate of non-synonymous substitutions. The evolutionary pattern shown by simian COB is very similar in magnitude and species range as that found previously by Adkins and Honeycutt (1994) for the cytochrome c oxidase subunit II (COII) gene. Investigation of the evolutionary rate of the cytochrome c oxidase subunit I (COI) gene also revealed that this gene had undergone a small evolutionary rate change similar to that of COII and COB. A survey of the evolutionary rates of all mitochondrial genes between apes and other non-primate mammals showed that in addition to the COB, COI and COII genes, the NADH dehydrogenase subunit 1 and 5 (ND1 and 5) and ATP synthase subunit 8 (ATP8) genes have also undergone non-synonymous rate increases in apes. Potentially, an episode of adaptive evolution of the simian primate ETC may encompass more than just complexes III and IV.

Presuming adaptive evolution among the genes of the mitochondrial genome of simian primates has taken place, then it would be of great interest to identify its cause. Co-evolution has been proposed previously as an explanation for the coordinated increase in evolutionary rates of cytochrome c and COII (Cann *et al* 1984), and this may also be the

case for COB and COI. In mammals, the only sphere where proteins encoded by both the nuclear and mitochondrial genomes come in to direct functional contact is the mitochondrial ETC, and this interaction between proteins from different genomes may have allowed these proteins to evolve in ways not observed elsewhere. In most other cases of adaptive evolution referred to in this thesis, adaptive change has occurred after functional constraint has been relaxed following gene duplication. This has not been the case for these mitochondrial genes. If adaptive evolution is indeed what has happened to the COI, COII and COB genes, then this is an unusual circumstance. Potentially, interaction of nuclear- and mitochondrial- encoded proteins could have a similar effect in terms of relaxing functional constraint as a gene duplication event does.

If adaptation of these mitochondrial genes has been the result of selection for a new or altered function of the ETC in simian primates, what could it have been, and what could have been the outcome? If a change to the ETC of simian primates was naturally selected, then the change must have either improved the efficiency or the output of electron transport. Potentially, an improvement to the ETC could either have provided the organism with greater efficiency in deriving energy from food, or have allowed the ETC to operate at higher capacity. Simian primates are unique among mammals in that they have overly-large brains for their body size, and an improvement to their ETC may have provided the extra energy required to power them. For mammals in general, when body weight is compared with brain weight a linear relationship is found (see Harvey and Bennett 1983, and references therein). The brain is an energetically expensive organ for the body to maintain, and brain weight is believed to be dictated by body size, as a body of sufficient size is required to provide the energy to power the brain. Other organs are also known to consume a large amount of the body's energy, and the current consensus is that simian primates make up the "energy-deficit" incurred due to their overly-large brains by having smaller guts (see Gibbons 1998). However, studies have shown that simian primates also have a high basal metabolic rate (measured by oxygen consumption) compared to other mammals, and it has been argued that simian primates support their bigger brains in a smaller body by generating more energy per unit of body weight than other mammals (reviewed in Armstrong 1990). It has been demonstrated that when mammalian body

mass is adjusted by a factor proportionate to their basal metabolic rate, simians have brain to body ratios similar to other mammals (Armstrong 1983; Armstrong 1985). This finding has been disputed (Harvey and Bennett 1983; Harvey and Krebs 1990), but it is remarkable that the species range of primates with overly-large brains and higher rates of respiration is the same as that identified here as having undergone a potential episode of adaptive evolution to the ETC. If an adaptive change to the ETC of simian primates did indeed improve the energy output of mitochondria sufficiently for simian primates to support bigger brains, this would be one of the defining changes in the evolution of these species. However, without hard biochemical evidence as to the relative catalytic efficiencies of simian and other primate ETC's it is only possible to speculate as to the cause of such a selective episode.

While it is apparent that three mitochondrial genes (COI, COII and COB) have elevated rates of non-synonymous substitution in simian primates (although see below), it does not automatically follow that each of these genes has been directly involved in a functional modification of the ETC. If large evolutionary change did occur to one protein subunit of a large multi-subunit complex such as cytochrome c oxidase, it would be likely that other proteins in the complex may change slightly to maintain their functional contact with this protein, or may change because certain functional constraints have been removed. Potentially, the non-synonymous change seen for the COI gene may just be compensation in response to modification of the COII protein. Each of the mitochondrial genes that show elevated evolutionary rates in simian primates demonstrate a higher rate of change not only after the point where an episode of adaptive evolution has been proposed to have occurred (after the divergence of simian primates from tarsiers), but also well after the divergences of the ape, Old World monkey and New World monkey species. Why this increased rate of evolution should be sustained for so long is not immediately obvious. What is proposed here is that because the COI, COII and COB proteins are part of a highly integrated biochemical system, it may have required a great many amino acid substitutions to completely optimise the interactions of these and other nuclear encoded proteins after an initial biochemical change occurred to key protein subunits. Fitch and Markowitz (1970) proposed the concept of the covarion, a group of COncomitantly VARIable codONs,

to explain why sequences of distantly related taxa can differ to the degree they do, while only a small number of positions are free to vary in any given lineage (also see Fitch 1971). The changes to the COI, COII and COB genes that occurred well after the proposed episode of adaptation may be the result of these proteins having changed covarion. Potentially, these proteins may have changed covarion more than once. After an episode of adaptive evolution to one or more subunits of the ETC, the covarions of a number of other proteins may have changed, and after these proteins changed covarion further rounds of compensating covarion change (probably of decreasing magnitude) could have taken place. Multiple rounds of covarion change may have occurred as a new equilibrium between ETC subunits was reached, and hence sustained non-synonymous evolution would have taken place on simian primate lineages well after an initial adaptive event. This proposal would be difficult to test, but given a sufficiently large number of closely related primate sequences for genes such as COI, it would be possible to estimate ancestral sequences and hopefully determine whether covarions have changed frequently in certain lineages.

In the preceding discussion the inference has been made that the increase in non-synonymous substitution rate seen in the simian primate COB, COI and COII genes implies adaptive evolution. While the work presented in this thesis provides considerable evidence that this is indeed is what has occurred, it has not been entirely possible to determine whether saturation of synonymous sites has led to a misleading result. Many of the mitochondrial genes analysed in Chapter 4 appeared to show saturation at synonymous sites, although a lack of primate sequence information meant it was not possible to ascertain at what level of divergence this was so. Hence, comparisons between species for these genes at synonymous sites would not show evolutionary rate heterogeneity even if it were present. However, with the data collected for the COI gene it appeared that saturation had not occurred at the evolutionary distances present between the primates, and synonymous substitution did not show an increase in rate between these species.

While the aim of the work presented in Chapter 5 was quite different to the rest of this thesis, it was a direct application of the knowledge obtained from preceding chapters. Removal of base compositional and rate heterogeneities from the primate mitochondrial

dataset allowed unbiased estimation of divergence dates among primates and more distantly related mammals. The difficulties encountered in estimating these divergence dates were mostly related to the small size of the mitochondrial dataset used, and illustrates the point that the larger the sequence dataset used the more likely an accurate result will be achieved.

A number of future experimental directions are immediately suggested by the results of the work presented here. The most important would be comparative biochemical analyses of complexes III and cytochrome c oxidase of the ETC. Assays of mitochondrial respiration in simian primate and other mammalian species would provide solid experimental evidence of whether differences in efficiency or throughput of the simian ETC exist, and whether the genetic evidence for positive selection obtained from this work means anything at a phenotypic level. Structural modelling of the important amino acid changes identified particularly for the interaction of cytochrome c and COII would provide information pertaining to which changes might have been important in any modulation of function of these proteins. Site directed mutagenesis experiments could be used to investigate any structural findings made, and it would be interesting to determine if functional differences from a simian ETC could be transplanted in non-simian ETCs. Mutagenesis of mitochondrial encoded proteins is a daunting experimental challenge, and this work could be conducted more fruitfully in the model system represented by the bacteria *Parracoccus dentrificans* and its crystallised three subunit cytochrome c oxidase (Iwata *et al* 1995).

The analysis of evolutionary rates of all mitochondrial protein-coding genes presented in Chapter 4 could be refined and expanded with sequence information provided by more complete mitochondrial genomes from primates. As was found for the COI gene, small evolutionary rate accelerations were more easy to detect over shorter periods of evolutionary time. With more mitochondrial sequence information, additional genes may be identified as having undergone small evolutionary rate accelerations. However, the benefit of this would not be as great as the identification of nuclear encoded genes such, as cytochrome c, that may be involved in the same phenomenon. As suggested in previous chapters, the Rieske iron-sulphur protein and cytochrome c_1 in complex III are

likely candidates due to their functional interaction with COB. Nuclear proteins in the cytochrome c oxidase complex may also be involved. Information on this matter from nuclear genes would not only aid understanding of the biochemical changes that may have taken place in the ETC of simian primates, but could provide insight into the evolutionary significance of interactions of proteins encoded by the nuclear and mitochondrial genomes.

To conclude, this work has shown the power that sequence-based genetic techniques have for the detection of adaptive evolution. Here, rate variations involving a number of simian primate ETC proteins have been uncovered which allude to an important evolutionary event in the history of these species. Experimentation on the mitochondria of multi-cellular eukaryotic organisms is difficult, and evolutionary differences in the ETC of primates can probably not be fully explored *in vivo*. However, genetic analysis has provided a starting point from which in silico structural modelling and *in vitro* mutagenesis can be conducted. While the analysis in this dissertation has been restricted to just one system, endless opportunities for the application of the methodologies applied here exist. An inescapable feature of genes that have been by positively selected is that something about them has made them more useful than other variants. Hence, positive selection is a marker of biological excellence and novelty. The potential for the use of positive selection to detect important biological phenomena is huge.

References

Adachi, J. and Hasegawa, M. (1995) Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J Mol Evol 40:622-628.

Adkins R. M., Honeycutt R. L. and Disotell, T. R. (1996) Evolution of eutherian cytochrome c oxidase subunit II: heterogeneous rates of protein evolution and altered interaction with cytochrome c. Mol Biol Evol 13:1393-1404.

Adkins, R. M. and Honeycutt, R. L. (1994) Evolution of the primate cytochrome c oxidase subunit II gene. J Mol Evol 38:215-231.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1989) *Molecular biology of the cell* Garland Publishing, New York.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. and Young, I. G. (1981) Sequence and organisation of the human mitochondrial genome. Nature 290:457-465.

Anderson, S., de Bruijn, M. H., Coulson, A. R., Eperon, I. C., Sanger, F. and Young, I. G. (1982) Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. J Mol Biol. 156:683-717.

Armstrong, E. (1983) Relative brain size and metabolism in mammals. Science 220:1302-1304.

Armstrong, E. (1985) Relative brain size in monkeys and prosimians. Am J Phys Anth 66:263-273.

Armstrong, E. (1990) Brains, bodies and metabolism. Brain Behav Evol 36:166-176.

Arnason, U., Bodin, K., Gullberg, A., Ledje, C. and Mouchaty, S. (1995) A molecular view of pinniped relationships with particular emphasis on the true seals. J Mol Evol 40:78-85.

Arnason, U., Gullberg, A. and Widegren, B. (1991) The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balenoptera physalus*. J Mol Evol 33:556-568.

Arnason, U., Gullberg, A. and Xu, X. (1996) A complete mitochondrial DNA molecule of the white-handed gibbon, *Hylobates lar*, and comparison among individual mitochondrial genes of all hominoid genera. Hereditas 124:185-189.

Arnason, U., Gullberg, A., Janke, A. and Xu, X. (1996) Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. J Mol Evol 43:650-661.

Arnason, U., Gullberg, A., Xu, X. and Graur, D. (1996) The "*Phoca* standard": an external molecular reference for calibrating recent evolutionary divergences. J Mol Evol 43:41-45.

Asenjo, A. B., Rim, J. and Oprian, D. D. (1994) Molecular determinants of human red/green color discrimination. Neuron 12:1131-1138.

Attardi, G. (1985) Animal mitochondrial DNA: an extreme example of genetic economy. Int Rev Cytol 93:93-145.

Avise, J. C. (1986) Mitochondrial DNA and evolutionary genetics of higher animals. Phil Trans R Soc Lond (Biol) 312:325-342.

Baba, M. L., Darga, L. L., Goodman, M. and Czelusniak, J. (1981) Evolution of cytochrome c investigated by the maximum parsimony method. J Mol Evol 17:197-213.

Bauer, C. and Jelkmann, W. (1977) Carbon-dioxide governs the oxygen affinity of crocodile blood. Nature 269:825-827.

Bauer, C., Forster, M., Gros, G., Mosca, A., Perella, M., Rollema, H. S. and Vogel, D. (1981) Analysis of bicarbonate binding to crocodile hemoglobin. J Biol Chem 256:8429-8435.

Bibb, M. J., Van Etten, R. A., Wright, C. T., Walberg, M. W. and Clayton, D. A. (1981) Sequence and gene organisation of mouse mitochondrial DNA. Cell 26:167-180.

Brand, M. D. and Murphy, M. P. (1987) Control of electron flux through the respiratory chain in mitochondria and cells. Biol Rev 62:141-193.

Brasseur, G., Saribas, A. S. and Daldal, F. (1996) A compilation of mutations located in the cytochrome b subunit of the bacterial and mitochondrial bc_1 complex. Biochim Biophys Acta 1275:61-69.

Brown, W. M. (1979) Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA 76:1962-1971.

Buffetaut, E. (1979) Evolution of the crocodilians. Sci Am 241:124-132.

Cann, R. L., Brown, W. M. and Wilson, A. C. (1984) Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. Genetics 106:479-499.

Capaldi, R. (1990) Structure and function of cytochrome c oxidase. Ann Rev Biochem 59:569-596.

Capaldi, R. A., Malatesta, F. and Darley-Usmar, V. M. (1983) Structure of cytochrome c oxidase. Biochim Biophys Acta 726:135-148.

Cavalier-Smith, T. (1987) The origin of eukaryotic and archaebacteria cells. Ann New York Acad Sci 503:17-54.

Chan, T., Lee, M. and Sakmar, T. P. (1992) Introduction of hydroxl-bearing amino acids causes bathochromatic spectral shifts in rhodopsin. J Biol Chem 267:9478-9480.

Chen, R., Greer, A. and Dean, A. M. (1995) A highly active decarboxylating dehydrogenase with rationally inverted co-enzyme specificity. Proc Natl Acad Sci USA 92:11666-11670.

Clayton, D. (1982) Replication of animal mitochondrial DNA. Cell 28:693-705.

Clementi, M. E., Condo, S. G., Castagnola, M. and Giardina, B. (1994) Hemoglobin function under extreme life conditions. Eur J Biochem 233:309-317.

Collura, R. V. and Stewart, C. B. (1995) Insertion and duplications of mtDNA in the nuclear genomes of old world monkeys and hominoids. Nature 378:485-489.

Dartnall, H. J. A., Bowmaker, J. K. and Mollon, J. D. (1983) Microspectrophotometry of human photoreceptors. In *Colour vision* (Eds. Mollon, J. D. and Sharpe, L. T.), Academic Press, New York, NY.

Dobson, D. E., Prager, E. M. and Wilson, A. C. (1984) Stomach lysozymes of ruminants. I. Distribution and catalytic properties. J Biol Chem 259:11607-11616.

Easteal, S. and Herbert, G. (1997) Molecular evidence from the nuclear genome for the time frame of human evolution. J Mol Evol 44:S121-S132.

Easteal, S., Collet, C. C. and Betty, D. J. (1995) *The mammalian molecular clock* Springer-Verlag, Austin, TX.

Endo, T., Ikeo, K. and Gojobori, T. (1996) Large-scale search for gene on which positive selection may operate. Mol Biol Evol 13:685-690.

Evans, M. J. and Scarpulla, R. C. (1988) The human somatic cytochrome c gene: two classes of processed pseudogenes demarcate a period of rapid molecular evolution. Proc Nat Acad Sci USA 85:9625-9629.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376.

Fitch, W. (1977) On the problem of discovering the most parsimonious tree. Am Nat 111:223-257.

Fitch, W. M. (1971) Rate of change of concomitantly variable codons. J Mol Evol 1:84-96.

Fitch, W. M. and Markowitz, E. (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-593.

Fitch, W. M., Leiter, J. M., Li, X. and Palese, P. (1991) Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. USA 88:4270-4274.

Fox, T. D. (1987) Natural variation in the genetic code. Ann Rev Genet 21:67-91.

Gadaleta, G., Pepe, G., De Candia, G., Quagliariello, C., Sbisa, E. and Saccone, C. (1989) The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates. J Mol Evol 28:497-516.

Gellissen, G. and Michaelis, G. (1987) Gene transfer: mitochondria to nucleus. Ann New York Acad Sci 503:391-401.

Gingerich, P. D. (1986) Temopral scaling of molecular evolution in primates and other mammals. Mol Biol Evol 3:205-221.

Gibbons, A. (1998) Solving the brain's energy crisis. Science 280:1345-1347.

Giles, R. E., Blanc, H., Cann, H. M. and Wallace, D. C. (1980) Maternal inheritance of human mitochondrial DNA. Proc. Natl. Acad. Sci. USA 77:6715-6719.

Golding, G. B. and Dean, A. M. (1998) The structural basis of molecular adaptation. Mol Biol Evol 15:355-369.

Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725-736.

Grivell, L. A. (1984) Mitochondrial DNA. Sci Am 248:60-73.

Groves, C. P. (1989) *A theory of human and primate evolution* Oxford University Press, Oxford.

Harvey, P. H. and Bennett, P. M. (1983) Brain size, energetics, ecology and life history patterns. Nature 306:314-315.

Harvey, P. H. and Krebs, J. R. (1990) Comparing brains. Science 249:140-146.

Hasegawa, M. (1990) Phylogeny and molecular evolution in primates. Jpn J Genet 65:243-265.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174.

Hasegawa, M., Kishino, H. and Yano, T. (1989) Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. J Hum Evol 18:461-476.

Hatefi, Y. (1985) The mitochondrial electron transport and oxidative phosphorylation system. Ann Rev Biochem 54:1015-1069.

Heibl, I., Braunitzer, G. and Schneeganss, D. (1987) The primary structures of the major and minor hemoglobin-components of the adult Andean goose (*Cloephaga melanoptera Anatidae*): the mutation Leu->Ser in position 55 of the b-chains. Biol Chem Hoppe Seyler 368:1559-1569.

Hillis, D. M., Marble, B. K. and Moritz, C. (1996) Applications of molecular systematics In *Molecular systematics* (2nd edition) Eds. Hillis D. M., Moritz, C. and Marble, B. K. Sinauer, Sunderland, MA.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. and Takahata, N. (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92:532-536.

Horai, S., Satta, Y., Hayasaka, K., Kondo, R., Inoue, T., Ishida, T., Hayashi, S. and Takahata, N. (1992) Man's place in Hominoidea revealed by mitochondrial DNA genealogy. J Mol Evol 35:32-43.

Hughes, A. L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170.

Hughes, A. L. and Yeager, M. (1997) Coordinated amino acid changes in the evolution of mammalian defensins. J Mol Evol 44:675-682.

Ina, Y. (1996) Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. J Genet 75:91-115.

Irwin, D. M. and Arnason, U. (1994) Cytochrome b gene of maritime mammals: phylogeny and evolution. J Mamm Evol 2:37-55.

Irwin, D. M., Kocher, T. D. and Wilson, A. C. (1991) Evolution of the cytochrome b gene of mammals. J Mol Evol 32:128-144.

Irwin, D. M., Prager, E. M. and Wilson, A. C. (1992) Evolutionary genetics of ruminant lysozymes. Animal Genetics 23:193-202.

Iwata, S., Ostermeler, C., Ludwig, B. and Michel, H. (1995) Structure at 2.8Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*. Nature 376:660-669.

Janke, A., Feldmaier-Fuchs, G., Thomas, W. K., von Haessler, A. and Pääbo, A. (1994) The marsupial mitochondrial genome and the evolution of placental mammals. Genetics 137:243-256.

Jermann, T. M., Opitz, J. G., Stackhouse, J. and Benner, S. A. (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 374:57-59.

Jermiin L. S., Olsen G. J., Mengersen K. L. and Easteal S. (1997) Majority-rule consensus of phylogenetic trees obtained by maximum-likelihood analysis. Mol Biol Evol 14:1296-1302.

Jermiin, L. S., Graur, D., Lowe, R. M. and Crozier, R. H. (1994) Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. J Mol Evol 39:160-173.

Jermiin, L. S., Wilson, S. R. and Easteal, S. (1998) Assessment of compositional stationarity in nucleotide and amino acid sequences. In review.

Jessen, T. H., Weber, R. E., Fermi, G., Tame, J. and Braunitzer, G. (1991) Adaptations of bird hemoglobins to high altitudes: demonstration of molecular mechanism by protein engineering. Proc Natl Acad Sci USA 88:6519-6522.

Jollès, P. and Jollès, J. (1984) What's new in lysozyme research? Always a model system, today as yesterday. Molec Cell Biochem 63:165-189.

Kimura, M. (1968) Evolutionary rate at the molecular level. Nature 217:624-626.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120.

Kimura, M. (1983) *The neutral theory of molecular evolution* Cambridge University Press, Cambridge.

Kimura, M. (1989) The neutral theory of molecular evolution and the world view of the neutralists. Genome 31:24-31.

Kishino, H. and Hasegawa, M. (1990) Converting distance to time: application to human evolution. Meth Enzymol 183:550-570.

Komiyama, N. H., Miyazaki, G., Tame, J. and Nagai, K. (1995) Transplanting a unique allosteric effect from crocodile into human haemoglobin. Nature 373:244-246.

Kornegay, J. R. (1996) Molecular genetics and evolution of stomach and nonstomach lysozymes in the hoatzin. J Mol Evol 42:676-684.

Kornegay, J. R., Schilling, J. W. and Wilson, A.C. (1994) Molecular adaptation of a leaf-eating bird: stomach lysozyme of the Hoatzin. Mol Biol Evol 11:921-928.

Kreitman,. M. and Akashi, H. (1995) Molecular evidence for natural selection. Ann Rev Ecol Syst 26:403-422.

Krettek, A., Gullberg, A. and Arnason, U. (1995) Sequence analysis of the complete mitochondrial DNA molecule of the hedgehog, *Erinaceus europaeus*, and the phylogenetic position of the *Lipotyphla*. J Mol Evol 41:952-957.

Kumar, S. and Hedges, S. B. (1998) A molecular timescale for vertebrate evolution. Nature 392:917-920.

Kumar, S., Tamura, K. and Nei, M. (1993) MEGA: Molecular Evolutionary Genetics Analysis., 1.01. The Pennsylvania State University.

Kunkel, T. A. and Loeb, L. A. (1981) Fidelity of mammalian DNA polymerases. Science 213:765-767.

Leciercq, F., Schenk, A. G., Braunitzer, G., Stangi, A. and Schrank, B. (1981) Direct reciprocal allosteric interaction of oxygen and hydrogen carbonate: sequence of the hemoglobins of the caiman (*Caiman crocodylus*), the Nile crocodile (*Crocodylus niloticus*) and the Mississippi crocodile (*Alligator mississippienis*). Hoppe-Seyler's Z. Physiol Chem 362:1151-1158.

Li, W-H. (1993) Unbiased estimation of the rates of synonymous and non-synonymous substitution. J Mol Evol 36: 96-99.

Li, W-H. (1997) *Fundamentals of Molecular Evolution*. Sinauer, Sunderland, MA.

Li, W-H., Wu, C-I. and Luo, C-C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150-174.

Linnane, A. W., Zhang, C., Baumer, A. and Nagley, P. (1992) Mitochondrial DNA mutation and the ageing process: bioenergy and pharmacological intervention. Mut Res 275:195-208.

Long, M. and Langley, C. H. (1993) Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science 260:91-95.

Lopez, J. V., Cevario, S. and O'Brien, S. J. (1996) Complete nucleotide sequences of the domestic cat (*Felis catus*) mitochondrial genome and transposed mtDNA repeat, *Numt*, in the nuclear genome. Genomics 33:229-246.

Ma, D-P., Zharkikh, A., Graur, D., VandeBerg, J. L. and Li, W-H. (1993) Structure and evolution of opossum, guinea pig and porcupine cytochrome b genes. J Mol Evol 36:327-334.

McDonald, J. H. and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652-654.

Messier, W. and Stewart, C.-B. (1997) Episodic adaptive evolution of primate lysozymes. Nature 385:151-154.

Metz, E. C. and Palumbi, S. R. (1996) Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol Biol Evol 13:397-406.

Miquel, J. (1992) An update on the mitochondrial-DNA mutation hypothesis of cell aging. Mut Res 275:209-216.

Mollon, J. D. (1991) The uses and evolutionary origins of primate-colour vision. In *Evolution of the eye and visual pigments*. (Eds. Cronly-Dillon, J. R. and Gregory, R. L.), CRC Press, Boca Raton, Fla.

Muse, S. V. and Gaut, B. S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715-724.

Muse, S. V. and Weir, B. S. (1992) Testing for equality of evolutionary rates. Genetics 132:269-276.

Nathans, J. (1990) Determinants of visual pigment absorbance: identification of the retinylidene Schiff's base counterion in bovine rhodopsin. Biochemistry 29:9746-9752.

Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426.

Nei, M., Zhang, J. and Yokoyama, S. (1997) Color vision of ancestral organisms of higher primates. Mol Biol Evol 14:611-618.

Nicholls, D. G. and Ferguson, S. J. (1992) *Bioenergetics 2* Academic Press, London.

Nicholls, D. G. and Locke, R. M. (1984) Thermogenic mechanisms in brown fat. Physiol Rev 64:1-64.

Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Olsen, G. J., Matsuda, H., Hagstrom, R. and Overbeek, R. (1994) fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comp App Biosci 10:41-48.

Osheroff, N., Speck, S. H., Margoliash, E., Veerman, E. C. I., Wilms, J., König, B. W. and Muijsers, A. O. (1983) The reaction of primate cytochromes c with cytochrome c oxidase. J Biol Chem 258:5731-5738.

Pamilo, P. and Bianchi, N. O. (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between genes. Mol Biol Evol 10: 271-281

Penefsky, H. S. and Cross, R. L. (1991) Structure and mechanism of F_0F_1-type ATP synthases and ATPases. Adv Enzymol 64:173-214.

Penny, D., Murray-McIntosh, R. P. and Hendy, M. D. (1998) Estimating times of divergence with a change of rate: the orangutan/African ape divergence. Mol Biol Evol 15:608-610.

Perutz, M. F. (1983) Species adaptation in a protein molecule. Mol Biol Evol 1:1-28.

Rambaut, A. and Bromham, L. (1998) Estimating divergence dates from molecular sequences. Mol Biol Evol 15:442-448.

Ramharack, R. and Deeley, R. G. (1987) Structure and evolution of primate cytochrome c oxidase subunit II gene. J Biol Chem 262:14014-14021.

Rice, W. R. (1989) Analyzing tables of statitical tests. Evolution 43:223-225.

Ruvolo, M. (1997) Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol Biol Evol 14:248-265.

Saitou, N. and Nei, M. (1987) The neighbour joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Sakmar, T. P., Franke, R. R. and Khorana, H. G. (1989) Glutamic acid-113 serves as the retinylidene Schiff base counterion in bovine rhodopsin. Proc Natl Acad Sci USA 86:8309-8313.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (2nd), Cold Spring Harbour Laboratory, Cold Spring Harbour, NY.

Schwartz, R. M. and Dayhoff, M. O. (1978) Origins of prokaryotes, eukarytotes, mitochondria and chloroplasts. Science 199:395-403.

Sears, C. (1994) Is it a cow? Is it a bird? New Scientist, 12 Nov: 35-39.

Seibert, S. A., Howell, C. Y., Hughes, M. K. and Hughes, H. A. (1995) Natural selection on the *gag*, *pol*, and *env* genes of human immunodeficiency virus 1 (HIV-1). Mol Biol Evol 12:803-813.

Shirai, T. and Go, M. (1997) Adaptive amino acid replacements accompanied by domain fusion in reverse transcriptase. J Mol Evol 44 (suppl 1):S155-S162.

Shoffner, J. M. and Wallace, D. C. (1992) Mitochondrial genetics: principles and practice. Am J. Hum Genet 51:1179-1186.

Singer, T. P., Kearney, E. B. and Kenney, W. C. (1973) Succinate dehydrogenase. Adv Enzymol 37:189-272.

Smith, C. A. B. (1986) Chi-squared tests with small numbers. Ann Hum Genet 50:163-167.

Stewart, C. B. and Wilson, A. C. (1987) Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. Cold Spring Harbor Symp Quant Biol 52:891-899.

Stewart, C. B., Schilling, J. W. and Wilson, A. C. (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401-404.

Strimmer, K. and von Haeseler, A. (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964-969.

Sutton, K. A. and Wilkinson, M. F. (1997) Rapid evolution of a homeodomain: evidence for positive selection. J Mol Evol 45:579-588.

Swanson, W. J. and Vaquier, V. D. (1995) Extraordinary divergence and positive Dawinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. Proc Natl Acad Sci USA 92:4957-4961.

Szkudlinski, M. W., Teh, N. G., Grossmann, M., Tropea, J. E. and Weintraub, B. D. (1996) Engineering human glycoprotein hormone superactive analogues. Nature Biotechnology 14:1257-1263.

Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512-526.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acid Res 22:4673-80.

Trumpower, B. L. (1990) The protonmotive Q cycle - energy transduction by coupling of proton translocation to electron transfer by the cytochrome bc_1 complex. J Biol Chem 265:11409-11412.

Trumpower, B. L. and Gennis, R. B. (1994) Energy transduction by cytochrome complexes in mitochondrial and bacterial respiration: the enzymology of coupling electron transfer reactions to transmembrane proton translocation. Ann Rev Biochem 63:675-716.

Tsaur, S.-C. and Wu, C.-I. (1997) Positive selection and the molecular evolution of a gene of male reproduction, *Acp26a* of *Drosophila*. Mol Biol Evol 14:544-549.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yoshikawa, S. (1995) Structures of metal sites of oxidised bovine heart cytochrome c oxidase at 2.8Å. Science 269:1069-1074.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yosikawa, S. (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8Å. Science 272:1136-1144.

Vacquier, V. D., Swanson, W. J. and Lee, Y.-H. (1997) Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? J Mol Evol 44 (suppl):S15-S22.

Vinogradov, A. D. (1993) Kinetics, control, and mechanism of ubiquinone reduction by the mammalian respiratory chain-linked NADH-ubiquinone reductase. J Bioenerg Biomemb 25:367-375.

Voet, D. and Voet, J. G. (1990) *Biochemistry* John Wiley and Sons, New York.

von Heijne, G. (1986) Why mitochondria need a genome. FEBS Lett 198:1-4.

Weiss, H., Friedrich, T., Hofhaus, G. and Preis, D. (1991) The respiratory-chain NADH dehydrogenase (complex I) of mitochondria. Eur J Biochem 197:563-576.

Whatley, J. M., John, P. and Whatley, F. R. (1979) From extracellular to intracellular: the establishment of mitochondria and chloroplasts. Proc R Soc Lond (Biol) 204:165-187.

Wilks, H. M., Hart, K. W., Feeney, R., Dunn, C. R., Muirhead, H., Chia, W. N., Barstow, D. A., Atkinson, T., Clarke, A. R. and Holbrook, J. J. (1988) A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. Science 242:1541-1544.

Witt, H., Malatesta, F., Nicoletti, F., Brunori, M. and Ludwig, B. (1998a) Cytochrome-c-binding site on cytochrome oxidase in *Paracoccus denitrificans*. Eur J Biochem 251:367-373.

Witt, H., Malatesta, F., Nicoletti, F., Brunori, M. and Ludwig, B. (1998b) Tryptophan 121 of subunit II is the entry electron site to cytochrome-c oxidase in *Paracoccus denitrificans*. J Biol Chem 273:5132-5136.

Wolf, M. J., Jermiin, L. S., Easteal, S., Kahn, S. and McKay, B. D. (1998) TrExML - a maximum likelihood program for exhaustive tree-space exploration. Submitted.

Wu, C.-I. and Li, W.-H. (1985) Evidence for higher rates of nucleotide substitution in rodents that in man. Proc Natl Acad Sci USA 82:1741-1745.

Wu, W., Goodman, M., Lomax, M. I. and Grossman, L. I. (1997) Molecular evolution of cytochrome c oxidase subunit IV: evidence for positive selection in simian primates. J Mol Evol 44:477-491.

Xia, D., Yu, C.-A., Kim, H., Xia, J.-Z., Kachurin, A. M., Zhang, L., Yu, L. and Deisenhofer, J. (1997) Crystal structure of the cytochrome bc_1 complex from bovine heart mitochondria. Science 277:60-66.

Xu, X. and Arnason, U. (1994) The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. Gene 148:357-362.

Xu, X., Gullberg, A. and Arnason, U. (1996a) The complete mitochondrial DNA (mtDNA) of the donkey and mtDNA comparisons among four closely related mammalian species-pairs. J Mol Evol 43:438-446.

Xu, X., Janke, A. and Arnason, U. (1996b) The complete mitochondrial DNA sequence of the greater Indian rhinoceros, *Rhinoceros unicornis*, and the Phylogenetic relationship among *Carnivora*, *Perissodactyla*, and *Artiodactyla* (+ *Cetacea*). Mol Biol Evol 13:1167-1173.

Yamaguchi, Y. and Gojobori, T. (1997) Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc Natl Acad. Sci USA 94:1264-1269.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comp Appl Biosci 13:555-556.

Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573.

Yang, Z., Kumar, S. and Nei, M. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641-1650.

Yoder, A. D., Cartmill, M., Ruvolo, M., Smith, K. and Vilgalys, R. (1996) Ancient single origin for Malagasy primates. Proc Nat Acad Sci USA 93:5122-5126.

Yokoyama, R. and Yokoyama, S. (1990) Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus*, and human. Proc Nat Acad Sci USA 87:9315-9318.

Yokoyama, S. (1995) Amino acid replacements and wavelength absorption of visual pigments in vertebrates. Mol Biol Evol 12:53-61.

Yokoyama, S. (1997) Molecular genetic basis of adaptive selection: examples from color vision in vertebrates. Ann Rev Genet 31:315-336.

Zhang, J., Rosenberg, H. L. and Nei, M. (1998) Positive Darwinian selection after duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708-3713.